Market Making with Costly Monitoring: An Analysis of the SOES Controversy

Thierry Foucault[1] Ailsa Röell[2] Patrik Sandås[3]

Current Draft: April 2000

[1]Department of Finance, HEC, School of Management, and CEPR, 1 rue de la Liberation, 78351 Jouy en Josas, France. Tel: (33) 1 39 67 94 11. Fax: (33) 1 39 67 70 85. E-mail:foucault@hec.fr

[2]Department of Economics, Princeton University, Princeton, NJ 08544. Tel: 609-248-4033. Fax: 609-258-64-19. E.mail:a.roell@princeton.edu.

[3]Finance Department, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104. Tel: 215-898-1697. Fax: 215-898-6200. E-mail:sandas@wharton.upenn.edu.

We are grateful to Dan Bernhardt, Maureen O'Hara (the Editor), Ingrid Werner and two anonymous referees for providing us with detailed comments. We also thank Bruno Biais, Hans Degryse, Frank Dejong, Burton Hollifield, Eugene Kandel, Ken Kavajecz, Gaëlle Lefol, Chester Spatt, Erik Theissen, and seminar participants at Berkeley, Carnegie-Mellon University, HEC, Leuven University, the Wharton Lunch Seminar, the 1999 WFA meetings at Santa Monica and the 1999 European Summer Symposium in Financial Markets at Gerzensee for useful comments. Financial support from the NASDAQ-AMEX Center for Financial Research and the Fondation HEC is gratefully acknowledged. All errors are ours.

Abstract

We develop a model of price formation in a dealership market where monitoring of the information flow requires costly effort. The result is imperfect monitoring, which creates profit opportunities for speculators who pick off 'stale quotes'. Externalities associated with monitoring give rise to multiple equilibria in which dealers earn strictly positive expected profits. We obtain various policy implications. A switch to automatic execution can improve or worsen (1) spreads and (2) price discovery depending on the specific equilibrium. A reduction in the minimum quoted depth tightens the spread but it reduces price efficiency. Our analysis is relevant for the SOES controversy given that speculators in our model behave as the real world SOES 'bandits'. Our model predicts that (1) SOES bandits should trade in stocks with small spreads and that (2) SOES bandit activity should widen the spread. We provide empirical evidence consistent with these predictions.

Keywords: Monitoring, Bid-Ask Spread, Automatic Execution, SOES trading.

1 Introduction

Nasdaq's Small Order Execution System (SOES) allows brokerage firms to automatically execute small orders at the best quotes posted by Nasdaq dealers. Participation in SOES is mandatory for all dealers, who must post firm quotes valid up to a maximum quantity, fixed by Nasdaq. Although it was intended for small retail customers, SOES mainly attracted professional day traders (labeled SOES 'bandits' by dealers). The bandits trade when they observe a shift in the value of the asset, either because they become aware of new public information before the dealers or because some dealers are slow to update their quotes.[1] This trading activity and its alleged impact on Nasdaq trading costs, liquidity, and volatility has been the subject of a long and heated policy debate.[2]

Harris and Schultz (1998) show that bandits on average make positive trading profits, at the expense of dealers. This observation is puzzling since bandits trade on information which is publicly available and pay commissions on their trades. Harris and Schultz (1998), p.61, suggest that imperfect monitoring by dealers is a potential explanation:

> *'The existence and profitability of SOES bandits raise new questions about the efficiency of different market structures. Bandits do not have any more information than the market makers that they trade against and in many cases they have less information. But bandits still make money. [...]. We believe the answer is that market makers are inherently less efficient at price discovery than are bandits. [...] bandits have a much greater incentives to concentrate on what they are doing, to follow stock prices closely, and to stay in front of their terminals than do market maker employees.'*

In this paper, we develop a model of market making with costly monitoring and show how this friction affects price formation. We distinguish between two forms of monitoring: (i) *market monitoring* and (ii) *quote monitoring*. Market monitoring entails monitoring the arrival of new information, e.g., public announcements, whereas quote monitoring is limited to monitoring quote updates. Market monitoring requires some effort. In contrast, quote monitoring does not require any effort because it can be automated.

[1] SOES day traders accounted for 83% of SOES share volume as of September 1995 according to the General Accounting Office (GAO) report on 'The effect of SOES on the Nasdaq Market'. Harris and Schultz (1997) provide a detailed description of their trading strategies.

[2] In a Washington Post article, on February 7, 1994, Joseph Hardiman, president of the National Securities Dealers Association said that 'The SOES activists were picking off market makers, who were slow to adjust. The losses to SOES activists made market makers gun shy, causing them to widen their price spreads.' In a testimony before the House Committee on Commerce, David Whitcomb, argued that 'Abolishing SOES would remove the 'market discipline', which keeps market makers on 'their toes' and causes prices to rapidly adjust when news occurs.' See the GAO report for a summary of the main arguments in the SOES controversy and important SOES-related events.

In our model, market makers post firm quotes and select how intensively they monitor information arrival. Since market monitoring is costly, they never monitor news continuously. Imperfect monitoring creates occasional profit opportunities due to 'stale' quotes. A second group of agents, referred to as speculators, seek to exploit these opportunities. They behave like the SOES bandits. When they observe new information or a quote revision indicating a change in the asset value, speculators 'pick off' dealers that fail to adjust their quotes. In equilibrium, speculators' expected profits are positive. The dealers' losses from trading with the speculators are offset by gains from trading with liquidity traders.

Our main results are:

1. Market monitoring by one dealer can generate either *a positive or a negative externality* for the other dealers. By monitoring quote updates, a dealer can free ride on the efforts that his competitors exert to monitor the market. This is the source of the positive externality. The negative externality stems from the fact that speculators can use quote updates to 'discover' stale quotes. The direction of the externality depends on the quickness with which dealers react to quote updates.

2. These externalities influence the dealers' bidding behavior. The positive externality induces dealers to match the best offers in the market rather than undercutting. This effect gives rise to multiple equilibria in which dealers earn strictly positive expected profits. In contrast the negative externality precludes the existence of equilibria in which more than one dealer can operate without incurring losses (a form of market breakdown).

3. Automatic execution of orders improves the speculators' ability to quickly respond to quote updates. For this reason, it has an impact on market quality. We show that a switch to automatic execution can increase or decrease spreads and the speed of price discovery depending upon which equilibrium is obtained.

Interestingly, the SOES's automatic execution feature has been a major bone of contention between bandits and Nasdaq dealers. Accordingly, Nasdaq has attempted to eliminate this feature several times. The policy debate has also focused on the effect of bandit activity on the spread. Dealers blamed bandits for being the cause of large spreads on Nasdaq and we obtain such an effect in our model. Despite its importance, there is no direct empirical evidence regarding this

question.[3] A difficulty is that the spread and the level of bandit activity are interdependent. In our framework an increase in the spread triggers the exit of some speculators. With the guidance of our model, we estimate a simultaneous-equations model which handles the interdependence between the spread and the level of bandit activity. Consistent with the predictions of our model, we find that widening the spread significantly lowers SOES bandit activity. We also find that stocks with a higher level of SOES bandit activity feature larger spreads. But the effect is surprisingly weak (significant at the 10% level), given dealers' insistance on the impact of SOES bandits on the spread.

Battalio, Hatch, and Jennings (1997) show that SOES bandits speed up the price discovery process and that SOES bandits' activity is positively related to price volatility. We obtain theoretical and empirical results consistent with their findings. Harris and Schultz (1997) analyze trading around a rule change that decreased the mandatory quoted depth from 1000 to 500 shares. They provide evidence suggesting that the rule change reduced the marker makers' losses to SOES bandits. Correspondingly, in our model, a decrease in the mandatory quoted depth leads to the entry of fewer speculators and thus tightens the spread. We also confirm empirically these predictions. We point out however that a decrease in the minimum quoted depth impairs price discovery.

Our model is most closely related to Copeland and Galai (1983), who analyze the free-trading option aspect of fixed quotes. We show how the free-trading option problem arises in equilibrium as a result of imperfect monitoring decisions by market makers. Kandel and Marx (1998) develop a theoretical model to study whether odd-eighth avoidance is a rational response by Nasdaq dealers to SOES bandits. In their model the profit opportunities of the SOES bandits are *implicitly* assumed to be due to imperfect monitoring by the dealers. Our contribution is to *explicitly* model imperfect market monitoring and to analyze its impact on the spread and the level of SOES activity in equilibrium. Kumar and Seppi (1994) model index arbitrage. Like the speculators in our model, arbitrageurs learn information from quote updates. But Kumar and Seppi (1994) assume that the arbitrageurs *always* observe quote updates faster than do dealers, which is not the case in our analysis. Furthermore the information structure is exogenous in their paper whereas it is endogenous here (through market monitoring). These two features are crucial for

[3] Harris and Schultz (1997) and Benston and Wood (1998) show that bandits trade ahead of short-term price changes and therefore inflict trading losses to dealers. This indirectly suggests that dealers' claim is correct since widening the spread is a way to cover trading losses.

our results.

The implications of our theoretical findings are discussed in the context of the SOES controversy. More generally our model provides insights regarding the implications of free-trading options for the design of trading systems. This is important since electronic trading systems increasingly features automatic execution.[4] This is the case of automated limit order markets. The NYSE also recently announced that it was considering implementing an automatic execution system for small orders.[5]

The rest of the paper is organized as follows. The general features of the model are presented in the next section. In Section 3, we show that market monitoring by one dealer can be a positive or a negative externality for the other dealers. In Section 4, equilibrium bidding strategies are analyzed. In Section 5, we study the effect of automatic execution on the spread and price discovery. Testable implications are derived in Section 6 and an empirical study of these implications is conducted in Section 7. The final section concludes. All proofs are in the Appendix.

2 The Model

2.1 Timing, Traders and Market Structure

There is a single risky asset with a liquidation value, $\tilde{V}$. At the beginning of the trading period, the expected liquidation value is v_0. There are three types of traders in this market: (i) $M \geq 2$ *dealers*, who post quotes, (ii) $N \geq 1$ *speculators* and (iii) *liquidity traders*, who submit market orders. Let $\mathcal{M}$ and $\mathcal{N}$ denote the set of all dealers and all speculators, respectively. All the traders are risk neutral.

A trading round comprises three stages, as described in Figure 1. In *the quoting stage*, dealers simultaneously determine their quoted spreads, $\{S_i\}_{i=1}^{i=M}$. The bid quote posted by dealer i is $b_i = v_0 - \frac{S_i}{2}$ and the ask quote is $a_i = v_0 + \frac{S_i}{2}$. Let $S_b = Min\{S_i\}_{i=1}^{i=M}$ be the *inside spread*, i.e., the smallest spread in the market. Dealers are required to honor their quotes for up to Q shares, the minimum quoted depth. For orders larger than Q, dealers can back away from their quotes.

[4] Stoll (1992) discusses the impact of automatic execution on the values of free trading options in limit order markets and in dealer markets. He points out that automatic execution can be detrimental to market quality because it increases the risk of being picked off for traders with stale quotes.

[5] See 'NYSE studying electronic trading system to fill small trades automatically', *Wall Street Journal*, November, 5, 1999.

In the second stage, *after* observing the quotes posted in the market, the dealers and the speculators choose their monitoring levels. The monitoring level chosen by a trader determines the probability that she is the first to discover a public announcement regarding the asset value (see below). We refer to the second stage as *the monitoring stage*.

Eventually, in *the trading stage*, one of the three following events occurs. With probability $\alpha < 1$, information arrives indicating an increase or a decrease in the asset value, with equal probabilities. In the case of an increase (decrease), the new expected value becomes: $v_1 = v_0 + \frac{\sigma}{2}$ ($v_1 = v_0 - \frac{\sigma}{2}$). In this case, speculators may submit market orders or dealers update their quotes (see below). With probability $(1 - \alpha)$, no information arrives. In this case, with probability $\beta > 0$, a buy or a sell market order is submitted by a liquidity trader, with equal probabilities. The expected size of the liquidity trader's order is δQ. Finally with probability $(1 - \beta)$, no order is submitted.

Incoming market orders are evenly split[6] among the dealers posting *the best quotes*. A liquidity trader places a single order and each dealer executes a fraction, $x^l(M_b) = 1/M_b$, of the order. A speculator can place one or more market orders depending on the number of dealers posting the inside spread. Specifically her trade size is: $Q^T(M_b) = x^s(M_b)M_bQ$, that is $x^s(M)$ is the fraction of the inside depth consumed by a speculator. Note that a dealer posting the inside spread trades $Q^T/M_b = x^s(M_b)Q$ shares against a speculator. Of course $x^s(1) = x^l(1) = 100\%$. For $M_b > 1$, we assume that $x^s(M_b)$ belongs to $[\frac{1}{M_b}, 1]$. The particular case in which a speculator places an order against each dealer corresponds to $x^s(M_b) = 100\%$.

The ratio of a speculator's order size to a liquidity trader's order size is:

$$o_{mix}(M) = \frac{x^s(M)}{x^l(M)\delta} = \frac{x^s(M)M}{\delta}$$

This ratio characterizes the mix of informed and non-informed orders in the order flow routed to a dealer. We call this ratio 'the order flow mix'. Obviously, the larger is the order flow mix, the greater is the adverse selection for a dealer.[7] Parameter δ controls variations in the order flow mix which are independent of the number of dealers. Parameter $x^s(M)$ controls variations

[6] Or market orders are routed to one dealer randomly chosen among the dealers posting the inside spread. The results are identical. Parameters $x^s(M)$ and $x^l(M)$ are interpreted as execution probabilities in this case.

[7] The order flow mix is also the proportion of informed trades to the proportion of non-informed trades. It plays the role of the proportion of 'bad' types to the proportion of 'good' types in the literature on adverse selection. Note that in general this ratio is assumed to be independent of the number of firms (here dealers) in the market.

in the order flow mix induced by a change in the number of dealers posting the inside spread. The case in which $x^s(M) = x^l(M)$ gives us the opportunity to analyze the effects of a change in the number of dealers at the inside, **holding the order flow mix constant**.

2.2 Market Monitoring and Quote Monitoring

Dealers and speculators become aware of new information by directly monitoring the information flow, an activity that we call *market monitoring*.[8] We model market monitoring as follows. Let $\lambda_i \in [0, +\infty)$ be *the monitoring level* of market-maker $i \in \mathcal{M}$ and let $\gamma_j \in [0, +\infty)$ be the monitoring level of speculator $j \in \mathcal{N}$. If new information arrives, the probability that a trader, say m, is first to observe the new information is denoted $Prob(f = m)$. This probability depends on the monitoring levels as follows:

$$Prob(f = i) \equiv P(\lambda_i) \equiv \frac{\lambda_i}{\lambda_i + \sum_{m \neq i} \lambda_m + \sum_j \gamma_j} \quad \forall i \in \mathcal{M}, \tag{1}$$

$$Prob(f = j) \equiv P(\gamma_j) \equiv \frac{\gamma_j}{\gamma_j + \sum_{k \neq j} \gamma_k + \sum_i \lambda_i} \quad \forall j \in \mathcal{N}. \tag{2}$$

Furthemore we set: $P(0) = 0$ and $P(+\infty) = 1$. A monitoring level equal to zero corresponds to the decision of *not monitoring the market at all*. Conversely, an infinite monitoring level corresponds to the decision of *continuously* monitoring the market. For any intermediate monitoring level there is some monitoring but it is imperfect. The probability that a trader will be first to react to new information increases with his (her) monitoring level and decreases with the aggregate monitoring level of the other traders.

Each trader must exert effort to support their chosen level of monitoring. The monetary disutility associated with this effort is captured by a cost function $\Psi(l)$, which is strictly increasing and *strictly convex* in the monitoring level l. We assume that[9]:

$$\Psi(l) = \frac{cl^2}{4}, \tag{3}$$

[8] Houtkin (1998) describes the trading strategies followed by SOES bandits. He provides a list of the events that they monitor. For example, business news, earnings announcements, price movements in related stocks, brokerage firms'upgrades and downgrades of stocks, announcements of economic indicators.

[9] Quadratic monitoring cost functions allow us to derive monitoring levels and equilibrium spreads in closed form. Our results only rely on the strict convexity of these functions. Furthermore the results are qualitatively similar when dealers on the one hand and speculators on the other hand have different monitoring costs functions.

The parameter $c > 0$ determines the scale of the monitoring cost for a given monitoring level. Speculators and dealers *simultaneously* choose their monitoring levels, after observing the inside spread.[10] We denote by $\lambda(S_b, M_b) = (\lambda_1(S_b, M_b), ..., \lambda_{M_b}(S_b, M_b))$, the vector of the dealers' monitoring levels. Dealers posting wider spreads than the inside spread optimally choose not to monitor since orders are only routed to the inside dealers. Correspondingly, $\gamma(S_b, M_b)$ denotes the vector of the speculators' monitoring levels.

Dealers and speculators also monitor quote updates (*quote monitoring*). Dealers use the information revealed by quote changes to update their offers and speculators use it to trade against dealers who are slow to adjust their quotes.[11] Quote monitoring does not contribute to price discovery but it enables traders to free ride on information production by dealers.

Traders can invest in software that alerts them to quote updates in different securities.[12] Therefore quote monitoring involves negligible variable costs. This means that the probability of being the first trader to react to a quote update is more likely to be determined by the trading technology used (or trading rules) than by the effort exerted. Accordingly we assume that when a dealer is first to update his quote, there is an exogenous probability Φ that one speculator reacts to this quote update before the $(M_b - 1)$ remaining dealers react. In this case, each speculator has an equal probability $(1/N)$ of being the speculator who first reacts. With probability $(1 - \Phi)$, the $(M_b - 1)$ remaining dealers update their quote before a speculator gets the chance to react to the initial quote update. Thus Φ is a measure of the relative advantage of speculators over dealers in quote monitoring (if $\Phi = 0$, dealers always react more quickly than speculators and vice versa if $\Phi = 1$).

[10] We have assumed that traders choose their monitoring level after observing the inside spread for two reasons. First, a trader's monitoring (effort) level is unobservable and therefore quotes cannot be made contingent on monitoring levels. Second, traders can adjust their effort, once the quotes have been posted. For this reason it is natural to assume that the monitoring levels are chosen after the quoting stage, i.e., can be contingent on the inside spread.

[11] Note that in our model quote revisions always imply changes in the asset value. In reality, quote revisions may occur for other reasons. A change in a dealer's inventory is an example. This means that a quote revision is a noisy signal of a shift in the asset value. However, the logic of the model applies insofar as quote revisions do contain information (occur because of shifts in the asset value).

[12] SOES day traders and dealers use softwares to detect stale quotes. See the GAO report.

2.3 Equilibrium

Assume that the inside spread is strictly lower than the size of the revision in the asset's expected value in case of information arrival, i.e., $S_b < \sigma$ (this will always be the case in equilibrium). Given our previous assumptions, the optimal course of action for the dealers and the speculators in the trading stage is as follows. If a dealer is first to observe the new information, he revises his quotes accordingly. If his competitors react to this quote update before the speculators, they revise their quotes as well. If a speculator is first to react to a quote update or to observe new information, she trades $Q^T(M_b)$ (in the direction of the quote revision). Tables 1 and 2 list the payoffs to the dealers and the speculators, for different decisions and outcomes in the monitoring and quoting stages.

We solve for the perfect equilibrium of the trading round, which is a set $\{S_b^*, M_b^*, \lambda^*(.,.), \gamma^*(.,.)\}$ such that **(i)** $\lambda^*(S_b, M_b)$ and $\gamma^*(S_b, M_b)$ form a Nash equilibrium of the monitoring stage for all possible outcomes of the quoting stage and, **(ii)** $\{S_b^*, M_b^*\}$ is a Nash equilibrium of the quoting stage. Although we always consider M as exogenous, the number of dealers (M_b) posting the inside spread is endogenous. Speculators' quote monitoring is meaningless when there is only one dealer at the inside. Hence the case in which $M_b = 1$ and $\Phi > 0$ is subsumed in the case $M_b = 1$ and $\Phi = 0$.

2.4 Discussion

Our trading game closely matches some of the key features of the Nasdaq's SOES trading system. The quantity, Q, is the mandatory quoted depth. The speculators can be thought of as the SOES bandits. In Nasdaq, dealers execute, at their posted quotes, orders that are larger than the minimum quoted depth. SOES bandits typically do not take part in these trades since they are negotiated by phone. This slows down the execution process and dealers can back away from their quote upon realizing that the counter-party is a bandit (See Harris and Schultz (1997) and Houtkin (1998)). Accordingly, the size of liquidity trades can be larger than Q (i.e. $\delta > 1$). Harris and Schultz (1998)'s findings suggest that bandits submit multiple orders. This is captured assuming that $x^s(M) > 1/M$. There are limitations on the number of positions initiated by individual bandits in Nasdaq. The effect of more stringent limitations can be analyzed by decreasing $x^s(M)$.

Our model features equilibria in which **only one** dealer can profitably post the inside spread.

In this case sidelined dealers are only exposed to speculators since liquidity traders are executed at the inside quotes. Sidelined dealers should therefore optimally widen their spreads so that picking off these dealers is not profitable. In order to account for this reaction within our *static* model of price competition, we simply assume that orders are not routed to the dealers who are not posting the inside spread.[13] This is in fact the case in SOES.

We assume that speculators unwind their positions at the mid-quote (v_1) subsequent to information arrival. This is a common assumption in models of trading with asymmetric information and it is particularly palatable for SOES bandits. Actually they frequently unload their positions on Selectnet or Instinet and trade within the quoted bid-ask spread. In fact Harris and Schultz (1998) find that when bandits lay off positions, they trade at or even above the spread mid-point in 90% of the cases. More generally, we could assume that speculators pay a fixed fraction τ of the spread when they close out their position (as in Kandel and Marx (1999)). They would then gain $(\sigma - (1+\tau)S)/2$ instead of $(\sigma - S)/2$ when they initiate a trade. This just scales up the effect of the spread on speculators' payoffs and this would not qualitatively affects our results. Hence we focus on $\tau = 0$ for brevity.

Finally note that the probability of a liquidity trade after an informational event is assumed to be zero. This assumption could easily be relaxed. Increasing the probability of a liquidity trade after an informational event essentially reduces the risk of being picked off for the dealers and is tantamount to a decrease in α.

3 Monitoring

3.1 Monitoring Externalities

In this section, we show that market monitoring by one dealer can generate a positive *or* a negative externality for the other dealers. These externalities play an important role in determining the quotes. We refer to them as monitoring externalities.

Consider one dealer, say i. There are two ways dealer i can be picked off when new information arrives. In the first case, a speculator reacts first to the information. This event occurs with probability $Prob(f \in \mathcal{N})$. Using Equation (2), we obtain:

[13] Another possibility would be to explicitly model price revisions. The equilibria we describe are robust to the possibility of price revisions in the sense that no dealer would find optimal to revise his quotes if he was offered the possibility to do so (before information arrival of course).

$$Prob(f \in \mathcal{N}) = \frac{\gamma_A}{\lambda_A + \gamma_A}. \tag{4}$$

where $\lambda_A \equiv \sum_i \lambda_i$ (resp. $\gamma_A \equiv \sum_j \gamma_j$) is the aggregate monitoring level of the dealers (resp. speculators). In the second case, a dealer (different from dealer i) observes the arrival of information, updates his quote and a speculator is first to react to the quote update. The probability of this event is $\Phi Prob(f \in \mathcal{M}_b \backslash i)$. Using Equation (1), we obtain:

$$Prob(f \in \mathcal{M}_b \backslash i) = \frac{\sum_{m \neq i} \lambda_m}{\lambda_A + \gamma_A}. \tag{5}$$

Let $\Pi_d(\lambda_i, \lambda_{-i}, \gamma, M_b)$ be dealer i's expected profit for given levels of monitoring, λ_{-i} and γ, for the other dealers and the speculators respectively. Using the payoff table (Table 1), we obtain:

$$\begin{aligned} \Pi_d(\lambda_i, \lambda_{-i}, \gamma, M_b) &= -\alpha \left[x^s(M_b) Prob(f \in \mathcal{N}) + x^s(M_b - 1) \Phi Prob(f \in \mathcal{M}_b \backslash i) \right] \frac{(\sigma - S_b)Q}{2} \\ &\quad + \left[(1-\alpha)\beta x^l(M_b) \right] \frac{S_b \delta Q}{2} - \Psi(\lambda_i) \quad \forall M_b \geq 2. \end{aligned} \tag{6}$$

The first term, which is negative, represents dealer i's expected loss from the risk of being picked off by speculators. The second term is positive and corresponds to dealer i's expected profit from trading with a liquidity trader. The last term is the monitoring cost incurred by dealer i. The probability of being picked off for dealer i is affected by the monitoring levels chosen by the other dealers. Thus market monitoring by one dealer is an externality for the other dealers. We obtain the following result.

Proposition 1 *: Consider two dealers i and m who are posting the inside spread. There exists a constant $\bar{\Phi} = \frac{\gamma_A x^s(M_b)}{(\gamma_A + \lambda_i) x^s(M_b - 1)}$ such that:*

1. *If $\Phi \leq \bar{\Phi}$ then market monitoring by dealer m is a positive externality for dealer i, or, $\frac{\partial \Pi_d(\lambda_i, \lambda_{-i}, \gamma, M_b)}{\partial \lambda_m} \geq 0$.*

2. *If $\Phi > \bar{\Phi}$ then market monitoring by dealer m is a negative externality for dealer i, or, $\frac{\partial \Pi_d(\lambda_i, \lambda_{-i}, \gamma, M_b)}{\partial \lambda_m} < 0$.*

Note that $x^s(M_b) \leq x^s(M_b - 1)$ is a sufficient (but not necessary) condition for $\bar{\Phi}$ to be lower than one. The economic intuition for this important property of market monitoring is as follows.

An increase in market monitoring by dealer m increases the probability that this dealer will be first to observe new information. This indirectly benefits dealer i since a quote update by dealer m *signals* to dealer i that his own quotes are misaligned. Thus, the increase in market monitoring by dealer m reduces dealer i's probability of being picked off through speculators' market monitoring ($\frac{\partial Prob(f\in\mathcal{N})}{\partial\lambda_m} < 0$). This is the source of the positive externality. However, there is a second effect since speculators monitor quote updates to learn about stale quotes. Accordingly, an increase in market monitoring by dealer m results in a greater probability of being picked off through speculators' quote monitoring for dealer i since $\frac{\partial Prob(f\in\mathcal{M}_b\backslash i)}{\partial\lambda_m} > 0$. This is the source of the negative externality. If dealer i reacts quickly enough to dealer m's quote updates ($\Phi \leq \bar{\Phi}$), the reduction in the risk of being picked off through market monitoring is larger than the increase in the risk of being picked off through quote monitoring. If speculators are relatively quicker ($\Phi > \bar{\Phi}$), the reverse is true.

3.2 Equilibrium in the monitoring stage

Dealer i chooses the monitoring level which maximizes $\Pi_d(\lambda_i, \lambda_{-i}, \gamma, M_b)$. Using the expression for dealer i's expected profit, the first order condition is:

$$-\alpha\left[x^s(M_b)\frac{\partial Prob(f\in\mathcal{N})}{\partial\lambda_i} + x^s(M_b-1)\Phi\frac{\partial Prob(f\in\mathcal{M}_b\backslash i)}{\partial\lambda_i}\right]\frac{(\sigma-S_b)Q}{2} = \Psi'(\lambda_i).$$

The terms inside the brackets measure the marginal reduction in the probability of being picked off due to increased monitoring by dealer i. The first order condition sets the marginal benefit of monitoring equal to the marginal cost. Using Equations (4) and (5), we rewrite the first order condition as:

$$\frac{\alpha Q^T(M_b)(\sigma-S_b)}{2M_b(\lambda_A+\gamma_A)^2}\left[\gamma_A + (\frac{x^s(M_b-1)}{x^s(M_b)})\Phi\sum_{m\neq i}\lambda_m\right] = \Psi'(\lambda_i), \tag{7}$$

The second order condition is satisfied if $S_b < \sigma$, which will be the case in equilibrium.

Let $\Pi_s(\gamma_j, \lambda, \gamma_{-j})$ be the expected profit for speculator j. Using Table 2, we obtain:

$$\Pi_s(\gamma_j, \lambda, \gamma_{-j}) = \frac{\alpha M_b Q(\sigma-S_b)}{2}\left[x^s(M_b)Prob(f=j) + \Phi x^s(M_b-1)\frac{Prob(f\in\mathcal{M}_b)}{N}\right] - \Psi(\gamma_j), \tag{8}$$

In the case of a change in the asset value, a profit opportunity arises because the dealers' quotes are temporarily mispriced. A speculator can capture this profit opportunity in two different ways: either (i) she is the first to react to the public announcement of a change in the asset value or (ii) she is the first to react to the quote update of a dealer. The term in bracket is the sum of

the probabilities of these two events, adjusted for the change in the expected total trade size for the speculator. Using Equation (1), we obtain:

$$Prob(f \in \mathcal{M}_b) = \frac{\lambda_A}{\lambda_A + \gamma_A}.$$

Speculator j chooses the monitoring level that maximizes $\Pi_s(\gamma_j, \lambda, \gamma_{-j})$. This implies setting the marginal benefit of monitoring equal to the marginal cost:

$$\frac{\alpha Q^T(M_b)(\sigma - S_b)}{2(\lambda_A + \gamma_A)^2} \left[\lambda_A (\frac{N - \Phi'}{N}) + \sum_{s \neq j} \gamma_s \right] = \Psi_s^{'}(\gamma_j), \tag{9}$$

where $\Phi' = \Phi \frac{Q^T(M_b - 1)}{Q^T(M_b)}$. The second order condition is satisfied if $S_b < \sigma$. Thus, a Nash equilibrium of the monitoring stage is a pair of vectors $(\lambda^*(S_b, M_b), \gamma^*(S_b, M_b))$ that solves Equations (7) and (9). This equilibrium is *symmetric* if all the traders of a given type (e.g., all the dealers) choose the same monitoring level.

Lemma 1 *: If there exists a Nash equilibrium in the monitoring stage, it is symmetric.*

Let λ^* (resp. γ^*) be the monitoring level chosen by each dealer (resp. speculator), in the symmetric equilibrium. The system of Equations ((7) and (9)) characterizing traders' best responses is then given by:

$$\frac{\alpha Q^T(M_b)(\sigma - S_b)}{M_b(M_b\lambda^* + N\gamma^*)^2} \left[N\gamma^* + \Phi' M_b \lambda^* \right] = c\lambda^*, \tag{10}$$

and

$$\frac{\alpha Q^T(M_b)(\sigma - S_b)}{(M_b\lambda^* + N\gamma^*)^2} \left[(\frac{N - \Phi'}{N}) M_b \lambda^* + (N-1)\gamma^* \right] = c\gamma^*. \tag{11}$$

Solving this system of equations yields the equilibrium monitoring levels. The solution is given in the next proposition.

Proposition 2 *: When M_b dealers post an inside spread $S_b < \sigma$, the equilibrium of the monitoring stage is unique and is characterized by the following monitoring levels for the speculators and the dealers (with $\Phi = 0$ if $M_b = 1$):*

$$\gamma^*(S_b, M_b) = \sqrt{\frac{\alpha N Q^T(M_b)(\sigma - S_b)}{c(\Upsilon + N)^2}}, \tag{12}$$

$$\lambda^*(S_b, M_b) = \frac{\Upsilon}{M_b} \gamma^*, \tag{13}$$

where $\Upsilon = \frac{N}{N-\Phi'}$. *For these monitoring levels, the expected profits of the speculators and the dealers are:*

$$\Pi_d(\lambda^*(S_b, M_b), \gamma^*(S_b, M_b), M_b) = \frac{Q}{2}\left[-\alpha x^s(M_b)(\sigma - S_b)C(M_b, \Phi) + (1-\alpha)x^l(M_b)\beta\delta S_b\right], \quad (14)$$

$$with \quad C(M_b, \Phi) \equiv \underbrace{\frac{N}{N+1-\Phi'}}_{Probability\ \ of\ \ being\ \ picked\ \ off} + \underbrace{\frac{N}{2M_b(N+1-\Phi')^2}}_{Monitoring\ \ Cost}, \quad (15)$$

and

$$\Pi_s(\lambda^*(S_b, M_b), \gamma^*(S_b, M_b), M_b) = \frac{\alpha Q^T(M_b)(\sigma - S_b)R'(N, \Phi)}{2N}, \quad (16)$$

$$with \quad R'(N, \Phi) \equiv \underbrace{\frac{N}{N+1-\Phi'}}_{Probability\ of\ Hitting\ Stale\ Quote} - \underbrace{\frac{(N-\Phi')^2}{2(N+1-\Phi')^2}}_{Monitoring\ \ Cost}. \quad (17)$$

The constant $C(M_b, \Phi)$ in dealers' expected profit function determines the size of the expected loss per unit of commited depth by a dealer. This is a measure of the (per unit) *cost of market making* for a dealer. This cost reflects the risk of being picked off and the monitoring cost borne by a dealer in equilibrium.

The proposition reveals several interesting properties of the monitoring strategies followed by the traders. Firstly speculators and dealers always put some effort in market monitoring ($\gamma^* > 0$ and $\lambda^* > 0$). In particular, it is never optimal for speculators to entirely base their trading strategies on dealers' quote updates.[14] Secondly, the monitoring level of both types of traders decreases with the size of the spread. When the dealers increase their spread, speculators monitor the market less intensively since the profit obtained by picking off dealers is lower. The dealers react by monitoring the market less intensively.

4 Spreads and Monitoring Externalities

The above results are all conditional on the spread. In this section we study the equilibrium spread. Our focus is on the relationships between the dealers' bidding strategies and monitoring externalities.

[14]This result is consistent with Harris and Schultz (1998) who, empirically, do not find a strong support for the view that bandits only trade after quote updates.

4.1 The Set of Equilibrium Spreads

Consider a situation in which *all* the dealers ($M_b = M$) post the inside spread S_b^*. This inside spread is a Nash equilibrium if no dealer has an incentive (i) to widen his spread or (ii) to improve upon the inside spread. The first condition requires that dealers do not expect to incur losses:

$$\Pi_d(\lambda^*(S_b^*, M), \gamma^*(S_b^*, M), M) \geq 0.$$

Let $\hat{S}(M, \Phi)$ be the spread such that this equation is binding. Using Equation (14), we get:

$$\hat{S}(M, \Phi) = \alpha\sigma\left(\frac{x^s(M)C(M, \Phi)}{\alpha x^s(M)C(M, \Phi) + (1-\alpha)x^l(M)\beta\delta}\right). \tag{18}$$

In equilibrium, the inside spread must be at least equal to $\hat{S}$ for the dealers to break even. Now suppose that a dealer improves slightly upon the inside spread. The expected profit for the dealer who undercuts is $\Pi_d(\lambda^*(S_b^*, 1), \gamma^*(S_b^*, 1), 1)$. In equilibrium, the dealer must be better off not undercutting. This requires that the profit earned by pooling on the inside spread with the other dealers is at least as high as the profit earned if the dealer undercuts:

$$G(S_b^*) = \Pi_d(\lambda^*(S_b^*, M), \gamma^*(S_b^*, M), M) - \Pi_d(\lambda(S_b^*, 1), \gamma^*(S_b^*, 1), 1) \geq 0. \tag{19}$$

Let $\Delta x^l(M)$ be the increase in a dealer' participation rate to uninformed trades when he undercuts ($\Delta x^l(M) = x^l(1) - x^l(M)$). We obtain:

$$G(S_b^*) = -\frac{\Delta x^l(M)Q}{2}\left[S_b^*(\alpha\bar{C}(M, \Phi) + (1-\alpha)\beta\delta) - \alpha\sigma\bar{C}(M, \Phi)\right],$$

where

$$\bar{C}(M, \Phi) = \frac{[C(1,0) - x^s(M)C(M, \Phi)]}{\Delta x^l(M)}. \tag{20}$$

Let $\bar{S}(M, \Phi)$ be the spread such that $G(\bar{S}(r)) = 0$, that is:

$$\bar{S}(M, \Phi) = \alpha\sigma\left(\frac{\bar{C}(M, \Phi)}{\alpha\bar{C}(M, \Phi) + (1-\alpha)\beta\delta}\right). \tag{21}$$

As $G(.)$ decreases with S_b^*, Equation (19) is satisfied if and only if $S_b^* \leq \bar{S}(M, \Phi)$. We conclude that S_b is an equilibrium spread if it belongs to $[\hat{S}(M, \Phi), \bar{S}(M, \Phi)]$. Using Equations (18) and (21), we obtain that the set of equilibrium spreads is non-empty **iff**:

$$\frac{x^s(M)}{x^l(M)}C(M, \Phi) \leq \bar{C}(M, \Phi)$$

or (using Equation (20) and dividing by δ):

$$\frac{x^s(M)}{x^l(M)}\frac{C(M, \Phi)}{\delta} \leq \frac{C(1,0)}{\delta} \tag{22}$$

This inequality has a simple interpretation. It states that an equilibrium in which all the dealers pool on the inside spread is viable iff the cost of market-making in this case is less than the cost of market of market-making when only one dealer posts the inside spread, accounting for a possible change in the order flow mix due to a change in the number of dealers at the inside. If this was not the case, a dealer at the inside would be better off undercutting.

We study in details the conditions under which Inequality (22) holds true below.[15] The next lemma will help the analysis.

Lemma 2 *: $\hat{S}(M,\Phi)$ increases with Φ whereas $\bar{S}(M,\Phi)$ decreases with Φ.*

An increase in speculators' relative advantage in quote monitoring enlarges the likelihood of a speculator intervention based on a quote update. This intensifies the adverse selection risk, which explains why the zero expected profit spread, $\hat{S}$, increases with Φ. Undercutting the inside spread is a way to prevent trades based on quote updates. This follows since orders are only routed to the dealer at the inside who therefore remains alone with an incentive to monitor. Accordingly a dealer has a greater incentive to improve upon the inside spread when Φ increases. This explains why such an increase tightens the largest possible spread ($\bar{S}$).

4.2 The Effect of Monitoring Externalities.

We now show that the positive externality associated with market monitoring helps dealers to earn strictly positive expected profits whereas the negative externality may result in a form of market breakdown. In order to better convey the intuition, we assume in this section that the order flow mix is independent of the number of dealers at the inside, i.e. $x^s(M) = x^s(M)$. Analysis of the case in which a change in the number of dealers affects the order flow mix is deferred to Section 4.3.

Note that for $\Phi = 0$, the cost of market-making decreases with the number of dealers posting the inside spread (see Equation (15)). This reflects the fact each dealer can free-ride on monitoring by a larger pool of dealers. He can therefore scale down his own monitoring without facing an increase in the risk being off. In fact pooling on the inside spread is a meachanism by which dealers (non-cooperatively) share the monitoring costs. Accordingly for $\Phi = 0$ and $x^s(M) = x^l(M)$, Inequality

[15] Note that the set of equilibrium spreads is independent of the absolute monitoring cost level, c. Thus the results hold even if the monitoring costs are small. The set of equilibrium spread is influenced by the relative monitoring cost levels between speculators and dealers which for simplicity has been normalized to 1.

(22) is always satisfied and we obtain the following result.

Proposition 3 *: In the absence of quote monitoring by speculators ($\Phi = 0$),*

1. *All the dealers post the inside spread in equilibrium (no sidelined dealers).*
2. *There is a multiplicity of equilibrium spreads: any spread $S_b \in [\hat{S}(M,0), \bar{S}(M,0)]$ is a Nash equilibrium. For all the equilibria in which the inside spread is strictly larger than $\hat{S}(M,0)$, the dealers earn strictly positive expected profits.*

In the quoting stage dealers compete in prices. The equilibrium does not necessarily feature zero expected profits for the dealers, however. Undercutting guarantees a larger share of the order flow to a dealer ('market share effect') but it prevents monitoring cost sharing ('cost sharing effect'). Actually, undercutting weakens the incentive to monitor for the dealers who are not at the inside and the dealer with price priority must support alone the burden of monitoring. For all spreads below $\bar{S}$, the resulting increase in the cost of market making is larger than the increase in revenue associated with a larger share of the order flow for a dealer who undercuts. Consequently, the positive externality associated with dealers' market monitoring helps to maintain spreads above the competitive level.

Lemma 3 *: If $\Phi > 0$, in equilibrium (a) either all the dealers post the inside spread ($M_b^* = M$) or (b) only one dealer posts the inside spread ($M_b^* = 1$).*

When $\Phi = 0$, an additional dealer posting the inside spread lowers the cost of market-making for all the dealers, even if initially only one dealer posts the inside spread. This explains why all the dealers pool on the inside spread in equilibrium. When $\Phi > 0$ and **only one** dealer posts the inside spread, there is another effect. An additional dealer at the inside enables the speculators to use the quote update of one dealer to pick off the other dealer. This effect triggers a discontinuous jump in the risk of being picked off. The cost of market making with two market makers may then be larger than with only one market-maker, despite the cost sharing effect. For this reason, matching the quotes of a single dealer can be sub-optimal and there may be equilibria with only one dealer at the inside. Not surprisingly, this depends on the speculators' ability to exploit the information contained in quote updates (the value of Φ) as shown in the next two propositions.

Proposition 4 *: There exists $\hat{\Phi}(M,N) \in (0,1)$ such that when $0 \leq \Phi \leq \hat{\Phi}(M,N)$:*

1. *All the dealers pool on the inside spread in equilibrium.*

2. *There is a multiplicity of equilibrium spreads: any spread* $S_b \in [\hat{S}(M,\Phi), \bar{S}(M,\Phi)]$ *is a Nash equilibrium. Dealers earn strictly positive expected profit when* $S_b > \hat{S}(M,\Phi)$.

When Φ is small, dealers react sufficiently quickly to quote updates (relative to speculators) for the cost sharing effect to still operate. When $\Phi > \hat{\Phi}$, speculators are relatively quicker than dealers in using the information contained in quote updates. Free-riding becomes too dangerous and sharing the monitoring burden is not an attractive option any more. In fact, dealers even find monitoring by their competitors undesirable since quote updates expose them dearly to speculators. An inside spread on which all the dealers pool would therefore be undercut. In this case Lemma 3 implies that the equilibrium must feature a single dealer posting the inside spread. It is described in the next proposition.

Proposition 5 *: When* $\hat{\Phi}(M,N) < \Phi \leq 1$*, the Nash equilibrium of the quoting stage is such that only one dealer (*$M_b^* = 1$*) posts the market spread which is* $S_b^* = \hat{S}(1,0)$*. The expected profit of the dealer posting the inside spread is zero.*

In order to prevent speculators from acquiring information through quote updates, dealers undercut each other until the point where a single dealer has no incentive to undercut the inside spread. For this reason, the dealer posting this spread just breaks even. The equilibrium in this case is also such that if another dealer were to match the inside spread, then the two dealers at the inside would incur losses. This is due to the discontinuous jump in the probability of being picked off we discussed previously. Note that this jump comes from the negative externality the two dealers inflict on each other (one dealer's quote update can be used to pick off the other one).

A too large advantage in quote monitoring for speculators over dealers dramatically reduces the supply of liquidity. No more than one dealer can operate in the market because two market makers or more posting the inside spread would incur losses. This sharp decline in liquidity when Φ becomes large is a form of *market breakdown*.[16] In particular, in Nasdaq, a listed stock must feature at least two market makers[17] but this not viable when $\Phi > \hat{\Phi}$ in our model. This observation provides some support to recent attempts by Nasdaq to reduce the advantage of

[16]We thank a referee for suggesting this interpretation.

[17]This is a requirement for continuous listing on Nasdaq SmallCap and NNM.

SOES bandits relative to dealers.[18] Interestingly, Nasdaq expressed the concern that trading by bandits may have caused a reduction in liquidity (a decline in the number of dealers) in stocks with high levels of SOES activity.

Figure 2 represents the set of equilibrium spreads as a function of Φ. When $\Phi \leq \hat{\Phi}$, there is a multiplicity of equilibrium spreads.[19] As usual in this situation, we must decide on which equilibria dealers are the most likely to coordinate (see Fudenberg and Tirole (1991)). We use the concept of *Pareto-Dominance* to select these equilibria. A Nash equilibrium is Pareto-Dominant if there is no other equilibrium that improves or leaves unchanged each dealer's expected profit.

Proposition 6 *: When there is a multiplicity of equilibria in the quoting stage, the unique Pareto-Dominant equilibrium is such that all dealers post the largest possible equilibrium spread,* $S_b^* = \bar{S}(M, \Phi)$.

The result is immediate since each dealer's expected profit increases with the inside spread (see Equation (14)). We will refer to the equilibrium in which dealer posts a spread equal to $\bar{S}(M, \Phi)$ as the Pareto-Dominant equilibrium and to the equilibrium in which dealers post a spread equal to $\hat{S}(M, \Phi)$ as the zero expected profit equilibrium. The Pareto-Dominant equilibrium is the preferred equilibrium for the dealer since it yields the largest expected profit. The zero expected profit equilibrium is obviously preferred by investors since execution costs are lower in this case.

4.3 The Effect of Multiple Orders by Speculators.

Dealers have an incentive to pool on the inside spread, even if it is above the competitive level, in order to share monitoring costs. This incentive might be defeated, however if such a pooling adversely affects the composition of the order flow. This can occur because an increase in the number of dealers at the inside gives the possibility to speculators to trade in larger sizes.[20] In

[18]For instance, in 1998, Nasdaq proposed a system in which dealers would be allowed to turn the system off momentarily while handling trades over the phone (see GAO report). Clearly such a possibility limits the risk of being picked off because of stale quotes for the dealers and corresponds to a decrease of Φ in our model.

[19]In addition to the equilibria described in the first part of Proposition 4, there is another equilibrium with $M_b^* = 1$ when $\Phi \leq \hat{\Phi}$ but Φ sufficiently greater than zero. In this equilibrium, the equilibrium spread is $S_b^* = \hat{S}(1, 0)$, which belongs to the set of possible equilibrium spreads described in the first part of Proposition 4. Kandel and Marx (1997) show that multiple equilibrium spreads can be obtained when a financial market features a positive tick size. Interestingly we obtain a multiplicity of equilibrium spreads even if the tick size is zero. A positive tick size would clearly not change this result but it could help dealers to coordinate on the Pareto-Dominant equilibrium.

[20]This means that the adverse selection problem is more intense in markets with a large number of dealers. In general, this effect is absent in models of trading with asymmetric information in the market microstructure literature. An exception is Dennert (1993). In our terminology, Dennert (1993) only analyzes the case in which

order to consider this possibility, we now allow for variations in the order flow mix when the number of dealers posting the inside spread changes, that is we set $x^s(M) > x^l(M)$. In this case:

$$\frac{o_{mix}(1) - o_{mix}(M)}{o_{mix}(M)} = 100(\frac{x^l(M)}{x^s(M)} - 1)\% < 0 \tag{23}$$

This quantity is negative and measures (in percent) the reduction in the ratio of a speculator's trade size to a liquidity trader's trade size (the order flow mix) when the number of dealers posting the inside spread is reduced from M dealers to 1 dealer.

Proposition 7 *: Suppose $\Phi = 0$. There exists $\bar{x}^s(M) \in (1/M, 1)$ such that if $1/M \leq x^s(M) \leq \bar{x}^s(M)$ then there is a multiplicity of equilibrium spreads as described in Proposition 3. ($\bar{x}^s(M)$ is given in the appendix).*

When $x^s(M) > x^l(M)$, undercutting the inside spread favourably rebalances the order flow that is routed to the dealer. Actually, a speculator's total trade size decreases whereas the liquidity trader's total trade size is unchanged. As a consequence the proportion of informed trades to the proportion of non-informed trades (the order flow mix) decreases. This effect adds to the market share effect described in the previous section: it reinforces the incentive to undercut. The cost sharing effect is still present, however: if he undercuts a dealer supports alone the burden of market monitoring. When the order flow mix is not too sensitive to the number of dealers at the inside ($x^s(M) \leq \bar{x}^s(M)$), the cost sharing effect dominates and spreads larger than the competitive spread can be sustained. When $x^s(M) > \bar{x}^s(M)$, the cost sharing effect is dominated and equilibria in which **all** the dealers pool on the inside spread do not exist. The question is then whether or not equilibria with less than M dealers but more than one dealer posting the inside spread can exist. In general, the answer to this question depends on the specific relationship between $x^s(M)$ and the number of dealers on which we made no assumptions. For $x^s(M) = 100\%, \quad \forall M$, we can provide an answer, however.

Proposition 8 *: Suppose $\Phi = 0$. If $x^s(M) = 100\%$, the only equilibrium is such that only one dealer posts the inside spread and he earns zero profit.*

In this case, undercutting the inside spread reduces the order flow mix by at least 100% (see Equation (23)), that is it dramatically reduces the proportion of informed trades relative to non

$x^s(M) = 100\%$. Furthermore market monitoring is not an issue in Dennert (1993) and therefore there is no cost sharing effect in his model.

informed trades. For this reason, a dealer is always better off undercutting the inside spread if two or more dealers post this spread, *even if these dealers just break even*. It follows that the equilibrium must feature a single dealer with zero profit (exactly as for $\Phi > \hat{\Phi}$). An additional dealer posting the equilibrium spread would trigger a discontinuous jump in the order flow mix (by at least 100%) so that no sidelined dealer find optimal to match the inside spread in this case. This suggests that unbridled trading by speculators can lead to a market breakdown (no equilibrium with more than one dealer posting the inside spread). This possibility vindicates Nasdaq' attempts to limit the number of orders submitted by SOES bandits. For instance, Nasdaq rules prohibit SOES traders for initiating more than one position in a given stock within five minutes.

Similar results are obtained when $\Phi > 0$. In particular it is clear that if $x^s(M) \leq \bar{x}^s(M)$ and if Φ is sufficiently close to zero, Inequality (22) is satisfied and there exist equilibrium spreads for which dealers capture strictly positive expected profits. Furthermore the effect of a change in Φ on the Pareto-Dominant equilibrium spread and on the zero expected profit spread are not affected by $x^s(M)$ (Lemma 2 is established for any value of $x^s(M)$). Overall equilibria with multiple dealers do not qualitatively differ when $x^s(M) = x^l(M)$ and when $x^s(M) > x^l(M)$. Hence, for simplicity, we assume from now on that $x^s(M) = x^l(M)$.

To sum up, in this section, we have shown how externalities associated with market monitoring influence the price formation process. The possibility for dealers to free-ride on market monitoring by other dealers reduces the incentive to undercut. Free-riding is more dangerous when speculators can pick off dealers based on quote updates, however. In fact, if speculators react sufficiently quickly to quote updates, a dealer can be hurt by other dealers' quote updates. This negative externality encourages dealers to undercut and makes it more difficult to sustain non competitive spreads. In the next section, we derive the implications of these results for market design.

5 Market Quality and Automatic Execution

Should dealers be protected against the automatic execution of stale quotes or not? This question has been central to the controversy between Nasdaq dealers and SOES bandits. Nasdaq dealers have argued that automatic execution made it easier for bandits to pick off dealers who were slow

to adjust their quotes. Accordingly dealers were obliged to widen their spreads.[21] In response, SOES bandits have argued that their presence has strengthened price competition among dealers. They also argued that they contributed to price discovery by forcing dealers to monitor more closely the stocks in which they were making the market.

Automatic execution certainly allows bandits to react more quickly when they observe stale quotes, that is it reduces dealers' advantage in quote monitoring. Accordingly, we can analyze the effect of suppressing automatic execution by comparing the case in which $\Phi = 0$ (automatic execution is prohibited) with the case in which $\Phi > 0$ (automatic execution is permitted).[22] Note that we already pointed out that granting a too large advantage to speculators over dealers in quote monitoring can result in a form of market breakdown. In the following analysis, we only focus on values of Φ such that this is not a concern ($\Phi \leq \hat{\Phi}$).[23]

Corollary 1 *:*

1. *When the equilibrium of the quoting stage is the Pareto-Dominant equilibrium, the inside spread is smaller when automatic execution is permitted.*

2. *When the equilibrium of the quoting stage is the zero expected profit equilibrium, the inside spread is larger when automatic execution is permitted.*

Consider Figure 2. If the dealers post the zero expected profit spread then the equilibrium spread is clearly larger when $\Phi > 0$. This reflects the fact that the adverse selection risk for the dealers is larger when speculators can use the information revealed by quote updates to pick off dealers. This supports dealers' argument that automatic execution increases the spread. On the other hand, if dealers post the Pareto-Dominant equilibrium spread, the conclusion is reversed: the equilibrium spread is smaller when $\Phi > 0$. Recall that the dealers' incentive to improve upon the inside spread is stronger when speculators can hit dealers who are slow to adjust their quotes.

[21] Nasdaq attempted to replace SOES with trading systems (N*Prove in 1994 and NAqcess in 1995) featuring delayed execution rather than automatic execution. The SEC never approved these systems.

[22] Nasdaq's Autoquote Policy prohibits software that would automatically update one market maker's quotes as a function of other market makers' quotes. By forcing a dealer to update his quotes manually when he receives an alert, this policy increases his reaction time. In our setting, we can also analyze the effect of suppressing the Autoquote policy by considering the impact of a decrease in Φ.

[23] This restriction does not affect our conclusions, however. That is the results hold also for the single dealer equilibrium which is obtained when $\Phi > \hat{\Phi}$.

This observation vindicates the SOES bandits' claim that they have increased price competition among dealers. This discussion uncovers one interesting effect of automatic execution: it makes free-riding market monitoring by other dealers more dangerous and for this reason it fosters price competition.

Corollary 2 *: The monitoring level chosen by a dealer in equilibrium is always larger when automatic execution is permitted, both in the zero expected profit and in the Pareto-Dominant equilibria.*

Automatic execution strengthens the dealers' incentive to be first to discover new information because it makes free-riding on other dealers' monitoring more dangerous. This effect is present whatever the nature of the equilibrium in the quoting stage and explains the result.[24]It vindicates the argument that automatic execution disciplines dealers and forces them to quickly reflect new information in their quotes.

The speed of price discovery is determined by the the *total* effort, $\lambda_A + \gamma_A$, exerted by all the traders in monitoring the market.[25] The following corollary compares the aggregate monitoring level when $\Phi = 0$ and when $\Phi > 0$.

Corollary 3 *:*

1. *When the equilibrium of the quoting stage is the Pareto-dominant equilibrium, the aggregate monitoring level, $\lambda_A^* + \gamma_A^*$, of all the traders is larger when automatic execution is permitted.*

2. *When the equilibrium of the quoting stage is the zero expected profit equilibrium, the aggregate monitoring level, $\lambda_A^* + \gamma_A^*$, of all the traders is smaller when automatic execution is permitted.*

Automatic execution may or may not improve price discovery. On the one hand, it strengthens dealers' incentives to monitor. On the other hand, it weakens speculators' incentive to monitor

[24]Automatic execution also has an indirect effect on dealers' market monitoring because it affects the equilibrium spread. The direction of the indirect effect depends on the equilibrium in the quoting stage. In the Pareto-Dominant equilibrium, automatic execution reduces the spread and in this way further enlarges dealers' market monitoring. In contrast, in the zero expected profit equilibrium, automatic execution widens the spread and in this way reduces dealers' need to monitor. Still this is insufficient for their equilibrium monitoring level to be smaller than when automatic execution is prohibited.

[25]In the model, the probability that one trader will discover whether or not an informational event has taken place is always equal to one. However, it is easy to modify the model in such a way that this probability is less than one, by adding a constant p in the denominators of $P(\lambda_i)$ and $P(\gamma_j)$. The probability that the informational event will *not* be discovered is then $\frac{p}{\lambda_A+\gamma_A+p}$. It decreases with $(\lambda_A + \gamma_A)$. Thus the speed of price discovery increases with $(\lambda_A + \gamma_A)$.

since they can use the costless information contained in quote updates to pick off dealers. In the zero expected profit equilibrium, this effect is reinforced by the fact that the spread is larger with automatic execution (γ^* decreases with the spread). It follows that in this case the aggregate monitoring is lower with automatic execution. On the contrary, in the Pareto Dominant equilibrium, the spread is smaller with automatic execution. In this case the increase in dealers' aggregate monitoring level is larger than the reduction in speculators' monitoring level and price discovery is improved.

6 Testable Implications

A major question in the SOES controversy is whether or not SOES bandits are responsible for wide spreads on Nasdaq, as claimed by dealers.[26] Our purpose is to study empirically this issue with the guidance of the model. We first consider the impact of an increase in the number of speculators on the equilibrium spread. Recall that the zero expected profit spread is:

$$\hat{S}(M, \Phi, N) = \alpha\sigma\left(\frac{C(M, \Phi)}{\alpha C(M, \Phi) + (1 - \alpha)\beta\delta}\right). \tag{24}$$

An increase in the number of speculators increases the risk of being picked off and results in a larger cost of market-making (larger C). The same effect holds for the Pareto-Dominant equilibrium. The next proposition follows.

Proposition 9 *: Other things equal, an increase in the number of speculators enlarges the equilibrium spread, both in the zero expected profit equilibrium and in the Pareto Dominant equilibrium.*

Therefore we predict that:

H.1: Other things equal, stocks with higher levels of SOES bandit activity have wider spreads.

Testing this prediction is not straightforward because the SOES bandit activity itself depends on the spread. Bandits should specialize in stocks with low spreads because their trading profits are larger for these stocks. In order to test our first prediction, we need to control for this effect which creates a negative correlation between the spread and the level of bandit activity.

[26] See also Grossman et al. (1995).

To this end, we extend the model assuming that each speculator bears a fixed entry cost, $K > 0$, that is sunk at the beginning of the trading game. This fixed cost represents, for instance, bandits' opportunity cost of time or the cost of freeing up capital and acquiring computer systems for trading. The number of speculators is then determined in such a way that a speculator's expected profit is just equal to the fixed cost.[27] For a given inside spread, S_b, a speculator's expected profit (see Proposition 2) net of the fixed cost K is:

$$\Pi_s(\lambda^*(S_b), \gamma^*(S_b), S_b, N) - K = \alpha Q(\sigma - S_b) \left[\frac{2N(N+1-\Phi) - (N-\Phi)^2)}{4N(N+1-\Phi)^2} \right] - K. \qquad (25)$$

This net expected profit decreases with the number of speculators and is negative when this number is large. For a given inside spread, the equilibrium number of speculators, $N^*(S_b)$, is such that the net expected profit is zero. Assume that $K \leq \Pi_s(\lambda^*(0), \gamma^*(0), 0, 1)$ (otherwise no trader would find it profitable to become a speculator even if the spread is zero) and that $\Phi \leq \hat{\Phi}(M, 1)$ (so that the equilibrium always features multiple dealers).

Proposition 10 *: Other things equal, an increase in the inside spread triggers a decrease in the number of speculators.*

An increase in the spread reduces a speculator's gain when she picks off a dealer. This explains the result. Hence our second prediction is that:

H.2: Other things equal, stocks with higher spreads have lower levels of SOES bandit activity.

Figure 3 depicts the effect of the number of speculators on the spread (the curve labelled S(N)) and the effect of the spread on the number of speculators (the curve labelled N(S)). The spread and the number of speculators are simultaneously determined and their equilibrium values (S^*, N^*) are at the intersection of these two curves. Consequently, we will test predictions H.1 and H.2 using a simultaneous-equations framework with the spread and the level of bandit activity as endogenous variables. In order to implement such a strategy we need to solve the associated identification problem and to control for variables which influence the spread and/or SOES bandit activity.

Consider the case in which dealers post the zero expected profit spread in equilibrium. Note that the average size of a liquidity trade (δ) and the number of dealers only appear in the spread equation (Equation (24)). This means that these variables do not *directly* determine the number

[27] There may not be an integer solution to the equality. In order to avoid this technical problem, we treat N as a real number, as it is usual in market entry analysis.

of speculators. The minimum quoted depth only appears in the speculator's expected profit equation which means that it does not directly affect the spread. Hence these variables (δ, M, Q) can be used as control variables and provide the exclusion restrictions that allow identification of our system of simultaneous-equations. The volatility of the asset (the size of the innovation) influences directly both the spread and the number of speculators used as a control variable. We obtain the following additional predictions.

Corollary 4 *:*

1. *For a given spread, an increase in the minimum quoted depth (Q) or an increase in volatility (σ) triggers an increase in the number of speculators.*

2. *For a given number of speculators, an increase in the average order size of liquidity trades (δ) or an increase in the number of dealers (M) triggers a decrease in the spread. An increase in volatility triggers an increase in the spread.*

Recall that the dealers posting the inside spread share the monitoring costs. It follows that the cost of market making and therefore the zero expected profit spread decrease with the number of dealers. The intuition for the other results is straightforward.

Note that the previous corollary signs the **direct** impact of a change in the exogenous parameters on the number of speculators holding the number of dealers **constant** and vice versa. In equilibrium, the spread and the number of speculators affect each other. Consider for instance an increase in the average size of liquidity trades. The direct effect, for a given number of speculators, is to decrease the spread (curve S(N) shifts downward). But the decrease in the spread triggers the entry of new speculators. This counterbalances the initial positive impact of a change in δ. Eventually the **total** impact of an increase in δ on the spread is smaller than the direct impact but it remains positive (see Figure 3 for an illustration). For all the exogenous parameters that affect only one of the endogenous variables, the direction of the total impact is the same as the direction of the direct impact (reported in Corollary 4).

The direct impact of volatility on both the spread and the number of speculators is positive. Hence, a priori, its total impact on each of these variables is ambiguous. In order to sign the total impact of volatility, we substitute the zero expected profit spread in Equation (25). This

yields:

$$(\frac{\alpha(1-\alpha)\beta\delta Q\sigma}{\alpha C(M,\Phi)+(1-\alpha)\beta\delta})\left[\frac{2N^*(N^*+1-\Phi)-(N^*-\Phi)^2)}{4N^*(N^*+1-\Phi)^2}\right]-K=0. \quad (26)$$

This equation implicitly defines the equilibrium number of speculators in term of the exogenous variables only and therefore allows to assess the total impact of these variables. When volatility increases, the R.H.S of this equation increases and more speculators find profitable to enter. The direct effect of an increase in volatility is to widen the spread (dealers' losses are larger when they are picked off). The positive impact of volatility on the number of speculators reinforces this effect (the spread increases with the number of speculators) and therefore the total impact of volatility on the spread is also positive. Eventually the total impact and the direct impact of volatility have the same direction.

Interestingly a change in the minimum quoted depth indirectly affects the spread because it has an impact on the number of speculators. The minimum quoted depth, Q, has been changed several times on Nasdaq; it was reduced from 1000 shares to 500 shares in January 1994, for most stocks, on a trial basis; it was restored to 1000 shares in March 1995 and eventually it has been reduced to 100 shares starting in January 1997. Nasdaq argued that the reduction of the minimum quoted depth would lessen SOES bandit activity and would narrow spreads. The next proposition concurs but it points out that a reduction in the minimum quoted depth adversely affects price discovery.

Proposition 11 *: In equilibrium:*

1. *When the minimum quoted depth decreases, the number of speculators decreases and the spread decreases.*
2. *When the minimum quoted depth decreases, the aggregate market monitoring* $(\lambda_A^* + \gamma_A^*)$ *decreases.*

A decrease in the minimum quoted depth induces the entry of fewer speculators since it reduces their expected profit in equilibrium (See Equation (26)). The risk of being picked off is lower and tighter spreads follow. In line with our prediction, Harris and Schultz (1997) find a decline in the number of trades initiated by SOES bandits after the reduction in the minimum quoted depth in 1994.[28] The reduction in the number of speculators implies that their aggregate monitoring

[28]Barclay *et al.* (1998) observe the same phenomenon after this quantity was reduced to 100 shares in 1997.

level decreases. Dealers choose to monitor less as well since the risk of being picked off is lower. Eventually price discovery is impaired.

The estimation procedure we use in the next section allows us to estimate both the direct and the total impacts of the exogenous variables on the spread and the number of speculators. We will therefore be able to check empirically whether spreads are positively related to the minimum quoted depth.

Corollary 4 and Proposition 11 are established considering the zero expected profit spread. It is straightforward to show that changes in parameters $\{Q, \delta, \sigma\}$ have similar effects in the Pareto Dominant equilibrium. The impact of a change in the number of dealers on the spread depends on the nature of the equilibrium, however. This is the next result.

Proposition 12 *: In the Pareto-Dominant equilibrium, for a given number of speculators, N, there exists $\Phi^*(M,N) \in (0, \hat{\Phi}(M,N))$ such that the inside spread decreases with the number of dealers if $\Phi \in [0, \Phi^*(M,N)]$ and increases with the number of dealers if $\Phi \in [\Phi^*, \hat{\Phi}(M,N)]$ (where $\Phi^*(M,N)$ is characterized in the proof).*

The extent of monitoring cost sharing increases with the number of dealers posting the inside spread. This makes undercutting less attractive when the number of dealers is large. On the other hand, each dealer executes a decreasing fraction of the order flow when the number of dealers increases. This effect encourages undercutting when the number of dealers is large. If Φ is sufficiently large, the first effect dominates and an increase in the number of dealers help dealers in sustaining larger non-competitive spreads.

Note that in the zero expected profit equilibrium, an increase in the number of dealers indirectly results in a larger number of speculators since it decreases the spread. In the Pareto-dominant equilibrium, this may or may not be the case since an increase in the number of dealers does not necessarily result in smaller spreads.

7 Empirical Analysis

7.1 Methodology

The actual number of SOES bandits is unobserved. A natural measure of their activity is the unconditional probability of observing a trade initiated by a bandit. In our model, this probability

is:

$$\alpha(Prob(f \in \mathcal{N}) + \Phi Prob(f \in \mathcal{M})) = \frac{\alpha N}{N + 1 - \Phi},$$

which is strictly increasing in the number of speculators N. The qualitative effects of a change in the exogenous parameters on the number of speculators and this probability are identical. Harris and Schultz (1997) show that SOES trades occuring in clusters (several maximum size SOES trades in rapid succession) are very likely to be initiated by bandits. We define a *cluster* as an uninterrupted sequence of three SOES orders of the maximum size, at the same price, within 30 seconds. Finally we use the probability of a SOES *cluster* as our proxy for the number of bandits. By focusing on clusters we avoid the risk of attributing SOES trades to bandits when in reality they are initiated by liquidity traders.

We estimate the following system of simultaneous-equations:

$$\begin{cases} soes_i = & a_1 + a_2 spr_i + a_3 vlty_i + a_4 maxQ_i + \epsilon_1 \\ spr_i = & b_1 + b_2 soes_i + b_3 vlty_i + b_4 ndlr_i + b_5 liqD_i + \epsilon_2, \end{cases} \tag{27}$$

where $i = 1, \ldots, N$ index the stocks and the variables in the equation system are: the probability of a SOES *cluster* (soes), the bid-ask spread (sprd), the volatility of the stock returns (vlty), the maximum quantity that can be traded in SOES (maxQ), the number of dealers in the stock (ndlr), and the average size of liquidity trades (liqD).

The first equation determines the probability of observing a SOES *cluster* as a function of the bid-ask spread, the volatility of the asset, and the maximum SOES quantity. The second equation determines the spread as a function of the probability of a SOES *cluster*, the volatility of the asset, the number of dealers, and the average size of liquidity trades. Our two main predictions are that the effect of the spread on the bandit activity is negative, $a_2 < 0$, and that the effect of the bandit activity on the spread is positive, $b_2 > 0$. Note however that our model has predictions (see Corollary 4) on the sign of the other independent variables as well.

7.2 Data

We use data provided by NASDAQ on transactions and dealer quotes for December 1996. Our sample consists of stocks with an average price above \$5 and a trading volume above four million shares for this month and it includes 310 stocks. Table 3 reports the mean, median, standard deviation, minimum, and maximum in the cross-section for the variables we use in the analysis. The first three rows report these statistics for the total number of SOES *clusters*, SOES trades,

and non-SOES trades. The bid-ask spread is measured as the time-weighted average inside spread. The standard deviation and the range suggest that there is substantial variation in both of these variables.

The volatility is measured by the standard deviation of the half-hour returns based on the mid-quotes, excluding overnight returns. The maximum SOES size is a discrete variable that is equal to 1000 (for 294 stocks), 500 (for 10 stocks), and 200 (for 6 stocks). The number of market makers for each stock is defined as the time-series average of the number of active market makers in the stock. The liquidity demand is measured by the average size of all trades, i.e., SOES trades as well as other trades, that were not part of a *cluster*. As expected these trades are on average larger than the maximum quantity that can be traded in the SOES system. The last two rows report statistics for the market capitalization and the average price for the sample. These two variables are likely to influence the bid-ask spread (see Harris (1994)) although they do not play a direct role in our model. We use them to improve the efficiency of our estimation.

In the actual estimation we use transformations of some of the variables discussed above. In the subsequent discussion our proxy for SOES bandit activity is defined as the logarithm of the odds ratio for clusters, i.e., $ln(\frac{p}{1-p})$, where p is the proportion of *clusters* among all trades.[29] We normalize the average trade size by the maximum SOES quantity so that the resulting variable, referred to as the liquidity demand below, corresponds to the δ in the model. Finally, we take the logarithm of the market capitalization and the average price.

Table 4 gives the correlation matrix for the variables that we use in the estimation. Notice that the cross-sectional correlation between the average bid-ask spread (sprd) and the proxy for soes bandit activity (soes) is -0.6835 in cross-section. Figure 4 plots the logarithm of the number of clusters (plus one) against the spread. It is evident from this graph that a regression of the spread on the number of clusters would generate a negative slope coefficient. But this would simply reflect the fact that more bandits are active in stocks with smaller spreads (Proposition 10). This does not rule out that an increase in bandit activity, *holding everything else equal*, leads to wider spreads as predicted by Proposition 9.

[29] There are a total of twelve stocks for which the total number of clusters is zero, eight of these stocks have a maximum SOES quantity of 1000 shares. To ensure that the log of the odds-ratio is always defined we add one to both the number of clusters and the total number of trades.

7.3 Empirical Results

Table 5 reports the parameter estimates and corresponding p-values for our model based on the equation system given in Equation (27).[30] The estimates for the endogenous variables provide support for the predictions of the model. The parameter estimate for the bid-ask spread in the SOES Equation is negative with a p-value less than 0.001. This means that an increase in the spread is an effective defense against trading by bandits. In fact we find some support to dealers' claim that bandits' attacks oblige them to widen their spreads: in the Spread Equation, the coefficient on bandit activity is positive. The effect of bandit on the spread is statistically weak however since the coefficient is significant only at the 10% level (p-value of 0.079). Possible explanations for this finding are provided in the next section.

The estimated coefficients for the volatility and the maximum SOES quantity in the SOES Equation are positive with p-values below 0.001. In the Spread Equation the coefficients on volatility and the number of dealers are positive (p-value-0.058) and negative (p-value < 0.001) respectively. All the estimates above have the predicted signs. The trade size does not appear to play an important role in determining the spread, the coefficient has a p-value of 0.453.

The estimated parameters in Table 5 measure the direct impact on the spread (resp. bandit activity) of one exogenous variable, holding other variables, including bandit activity (resp. the spread), constant. We also report the results of a linear regression of the endogenous variables on all the exogenous variables (so called 'reduced-form' regressions). The coefficients for these regressions can be interpreted as a measure of the total impact of a change in one exogenous variable on both the spread and bandit activity. The estimated parameters with p-values for these reduced form regressions are reported in Table 6.

Recall that the effect of a change in the minimum quoted depth on the spread is of particular interest. In Table 6, the coefficient on the maximum SOES quantity, in the Spread Equation, is positive (with a p-value of 0.058). Therefore, other things equal, stocks with a lower minimum quoted quantity have tighter spreads, as predicted by Proposition 11. The relatively low p-value may reflect the finding that SOES bandit activity has a moderately significant impact on the spread. Actually according to our model, the effect of the minimum quoted depth on the spread is indirect: an increase in this variable attracts bandit activity, which in turn tends to increase

[30] The system is estimated using three-stage least squares. The log of the market capitalization and the average price are added to the spread equation as additional control variables.

the spread.

In the zero expected profit equilibrium and in the Pareto-Dominant equilibrium (for Φ low enough), an increase in the number of dealers decreases the spread and therefore indirectly leads to more speculators. In line with this prediction, the coefficient on the number of dealers is positive in the SOES Equation and negative in the Spread Equation with p-values below 0.001.

In the SOES Equation we have a positive coefficient on volatility with a p-value of 0.259. Recall that total impact of volatility on bandit activity reflects a direct positive effect (confirmed in Table 5) and an indirect negative effect via the spread. The predicted sign of the total effect is positive as explained in the previous section. However, our empirical results suggest that the two effects essentially cancel so that volatility does not significantly affect the bandit activity. Interestingly, we observe the opposite phenomenon in the Spread Equation. Recall that the direct effect of volatility on the spread was positive with a p-value of 0.058. The total effect, which is reinforced by the positive effect that volatility has on bandit activity, is positive with a p-value of 0.012. All the total effects for the variables discussed above have the expected signs. Finally, the coefficients on liquidity demand are insignificant with p-values of 0.138 and 0.333 respectively.

7.4 Further Issues

While our results support the hypothesis that bandit activity has a positive effect on the spread, the evidence is statistically weak since the coefficient is significant only at the 10% level. This is somewhat surprising given that this issue has been a source of long conflict between the SOES bandits and the dealers. We ascribe this finding to several factors.

Given the endogeneity of the two key variables, the spread and the bandit activity, the validation of our predictions in the data hinges on finding good instruments. It is likely that a sample with greater variation in the key instruments, the maximum SOES quantity and the number of dealers, would produce more precise estimates. In the data that we use it is particularly hard to generate variation in the maximum SOES quantity within a sample of active stocks with a large number of market makers.

Previous research suggests that a substantial fraction of the SOES bandit activity is concentrated in a relatively small number of large active stocks (Harris and Schultz (1998), Kandel and Marx (1999)). Thus, it is possible that it would be easier to identify the effect of bandit activity on the spread in a time-series study of such stocks. However, it is worth noting that in order to

implement such a strategy a set of restrictions, different than the ones we used in a cross-section, has to be developed in order to test the two main predictions. For example, since there is little or no variation in a given stock's maximum SOES quantity over time an alternative instrument is needed to replace the maximum SOES quantity and to achieve identification. This seems like a natural direction for future research.

Finally, there are features of the trading organization in Nasdaq which make it difficult to measure the impact of bandit activity on the quoted spread. First, in our sample period, the minimum price increment was \$1/8. For some stocks, this may be larger than the compensation required by dealers for the risk of being picked off by bandits. In such a case, an increase in bandit activity will have no discernible impact on observed spreads even if it increases the cost of market-making. Second, on Nasdaq, traders have the possibility to negotiate with the dealers and many trades (especially large trades) receive price improvements (occurs at a price within the spread). In our model, dealers compensate the losses inflicted by bandits by posting larger spreads. In reality, they may decide to leave their quoted spread unchanged but to offer price improvements less frequently. For these reasons, it may be premature to conclude based on our results that the SOES bandit activity has a negligible impact on the cost of market-making and/or execution costs in Nasdaq.

8 Conclusion

We present a model of market-making with costly monitoring. We use the model to shed light on the main issues that arised in the SOES controversy. We find that when monitoring is costly, there is a strong incentive for dealers to pool on the inside spread in order to share monitoring costs. This incentive operates again price competition and leads to equilibria in which dealers earn strictly positive expected profits. This incentive is softened however when speculators can quickly use the information revealed by quote updates to pick off dealers who are slow to adjust their offers. For this reason automatic execution has an impact on market quality (spread and price efficiency). We also show that a reduction in the mandatory depth tightens the spread but that it weakens traders' incentive to monitor the information flow. Two main predictions of our analysis are that (i) stocks with larger spreads should feature lower levels of bandit activity, other things equal and that (ii) stocks with high levels of bandit activity should feature larger spreads, other things equal. Empirically we find a strong support for the first prediction and a moderate

support for the second.

Several extensions of our model of market-making with costly monitoring may be considered in future research. First our results do not rely on differences in monitoring skills among traders or agency issues. However, our model provides a natural framework to consider these possible explanations for bandits' profits. For instance, the monitoring decision can be interpreted as an unobservable (an non-contractible) choice of effort made by a trader who is an employee in a market-making firm. The issue arises as to which compensation scheme should be offered to this employee in order to maximize his incentive to monitor, more generally to provide high quality services in market-making. Second, in our model, it is always optimal for the market makers to adjust their quotes as they discover new information. This is consistent with the fact that dealers can not use SOES for proprietary trading.[31] In other market structures, the relevant choice may be to either update the quotes or to conceal the information and attempt to trade against other market participants.

[31] Before January 1997, market makers were not allowed to use SOES in any capacity. Since January 1997, they can use SOES on an agency basis only. See Smith et al. (1999). Note that even if dealers could trade on SOES they may choose not to do so for reputation reasons.

9 Appendix

Proof of Proposition 1

Using Equation (6):

$$\frac{\partial \Pi_d(\lambda_i, \lambda_{-i}, \gamma, M_b)}{\partial \lambda_m} = -\alpha \left[x^s(M_b) \frac{\partial Prob(f \in \mathcal{N})}{\partial \lambda_m} + x^s(M_b - 1)\Phi \frac{\partial Prob(f \in \mathcal{M}_b \backslash i)}{\partial \lambda_m} \right] \frac{(\sigma - S_b)Q}{2}$$

$$= \frac{\alpha}{(\lambda_A + \gamma_A)^2} \left[(1 - \frac{\Phi x^s(M_b - 1)}{x^s(M_b)})\gamma_A - (\frac{\Phi x^s(M_b - 1)}{x^s(M_b)})\lambda_i \right] \frac{(\sigma - S_b)x^s(M_b)Q}{2} \quad \forall m \neq i. \quad (28)$$

The R.H.S. of Equation (28) is positive **if and only if** the expression in square brackets is, i.e., as long as $(1 - \frac{\Phi x^s(M_b-1)}{x^s(M_b)})\gamma_A - (\frac{\Phi x^s(M_b-1)}{x^s(M_b)})\lambda_i \geq 0$. Hence $\bar{\Phi}$ follows directly.□

Proof of Lemma 1.

Suppose (to be contradicted) that there exists a Nash equilibrium in which some dealers do not choose the same monitoring levels. Consider two dealers i and i' such that $\lambda_i^* > \lambda_{i'}^*$. Using the fact that Equation (7) must hold for these two dealers, we obtain the following equality:

$$\frac{\alpha Q^T(M_b)(\sigma - S_b)}{2M_b(\lambda_A + \gamma_A)^2} \left[(\frac{\Phi x^s(M_b - 1)}{x^s(M_b)})(\lambda_{i'}^* - \lambda_i^*) \right] = \Psi^{'}(\lambda_i^*) - \Psi^{'}(\lambda_{i'}^*).$$

Since $\lambda_i^* > \lambda_{i'}^*$, the L.H.S of this inequality is strictly negative. But since $\Psi_d(.)$ is strictly convex, the R.H.S is strictly positive. This is impossible. This implies that in equilibrium all the dealers choose the same monitoring level. In the same way we can prove that in equilibrium all the speculators must choose the same monitoring level.□

Proof of Proposition 2.

Dividing Equation (10) by Equation (11), we find that λ^* and γ^* must satisfy:

$$\frac{N\gamma^* + \Phi' M_b \lambda^*}{\frac{(N-\Phi')}{N} M_b \lambda^* + (N-1)\gamma^*} = (\frac{M_b \lambda^*}{\gamma^*}).$$

This equation can be written as an equation with unknown $\Upsilon \equiv \frac{M_b \lambda^*}{\gamma^*}$:

$$\frac{N + \Phi' \Upsilon}{\frac{(N-\Phi')}{N}\Upsilon + (N-1)} = \Upsilon.$$

Since the monitoring levels must be positive, it must be the case that $\Upsilon \geq 0$. The previous equation has two solutions but only one is positive. This solution is: $\Upsilon = \frac{N}{(N-\Phi')}$. Substituting λ^* by $\frac{\Upsilon \gamma^*}{M_b}$ in Equation (11), we find that γ^* solves:

$$\frac{\alpha Q^T(\sigma - S_b)(\Upsilon(\frac{N-\Phi'}{N}) + (N-1))}{\gamma^*(\Upsilon + N)^2} = c\gamma^*.$$

There is a unique positive solution to this equation, which yields the closed form solution for γ^*. Since Υ and γ^* are uniquely defined, there is a unique Nash equilibrium in the monitoring stage. Substituting the expressions for λ^* and γ^* in Equations (4) and (5), we obtain that in equilibrium:

$$Prob(f \in \mathcal{N}) = \frac{N}{\Upsilon + N}, \tag{29}$$

and

$$Prob(f \in \mathcal{M}_b \backslash i) = \frac{(M_b - 1)\Upsilon}{M_b(\Upsilon + N)}. \tag{30}$$

Direct substitution of these probabilities in Equations (6) and (8) yield dealers' and speculators' expected profits.□

Proof of Lemma 2.

Using Equations (15) and (20), we get:

$$\frac{\partial C(M,\Phi)}{\partial \Phi} > 0 \qquad and \qquad \frac{\partial \bar{C}(M,\Phi)}{\partial \Phi} = -\frac{x^s(M)}{\Delta x^l(M)} \frac{\partial C(M,\Phi)}{\partial \Phi} < 0.$$

This proves the lemma since $\hat{S}$ ($\bar{S}$) increases with $C(M,\Phi)$ ($\bar{C}(M,\Phi)$).□

Proof of Proposition 3.

1st part. Suppose that the outcome of the quoting stage is $\{S_b, M_b\}$ with $M_b \geq 1$. Note that $C(M_b, 0)$ decreases with M_b. Therefore if:

$$\Pi_d(\lambda^*(S_b, M_b), \gamma^*(S_b, M_b), M_b) = \frac{x^s(M_b)Q}{2}[-\alpha(\sigma - S_b)C(M_b, 0) + (1-\alpha)\beta\delta S_b] \geq 0,$$

then

$$\Pi_d(\lambda^*(S_b, M_b{+}1), \gamma^*(S_b, M_b{+}1), M_b{+}1) = \frac{x^s(M_b+1)Q}{2}[-\alpha(\sigma{-}S_b)C(M_b{+}1, 0){+}(1{-}\alpha)\beta\delta S_b] > 0,$$

This means that a sidelined dealer is always better off matching the inside spread. Hence we can not construct an equilibrium in which a subset of dealers are sidelined when $\Phi = 0$.

2nd part. Since $C(M, 0)$ decreases with M, Inequality (22) is satisfied when $x^s(M) = x^l(M)$. The second part of the proposition is then immediate.□

Proof of Lemma 3.

For $M \geq 2$, $C(M,\Phi)$ decreases with M. Therefore we can proceed as in the proof of Proposition 3 (1st part) to show that there is no equilibrium in which a subset of two or more dealers post the spread and some dealers are sidelined. We can not discard the possibility that $C(1,0) < C(2,\Phi)$,

however. Therefore we can not exclude equilibria with only one dealer posting the inside spread. This proves the lemma. □

Proof of Proposition 4.

Let $\hat{\Phi}(M,N)$ be the value of Φ such that Inequality (22) is binding for $x^s(M) = x^l(M)$. It solves:

$$C(M,\Phi) - C(1,0) = (1+N-\Phi)\left[(1+N)(1-2\Phi)-\Phi\right] - \frac{(N+1)^2}{M} = 0.$$

The threshold $\hat{\Phi}(M,N)$ is:

$$\hat{\Phi}(M,N) = \frac{(1+N)(2+N)}{3+N}\left[1-\sqrt{1-\frac{(M-1)(3+N)}{M(2+N)^2}}\right].$$

Since $C(M,\Phi)$ increases with Φ, the proposition is proved. Notice that $\hat{\Phi}(.,.)$ increases with M and N. It is always lower than $1/2$.□

Proof of Proposition 5

When $\Phi > \hat{\Phi}$, Inequality (22) does not hold and there is no equilibrium in which all the dealers pool on the inside spread. If an equilibrium exists, it must therefore feature a single dealer (an implication of Lemma 3). We now prove that such an equilibrium exists.

For an equilibrium with only one dealer (say dealer m) posting the best offers to exist, three conditions must be satisfied. First dealer m should not be better off widening his spread. This requires $S_b^* \geq \hat{S}(1,0)$. This also requires that the spread posted by the sidelined dealers is just slightly greater than the inside spread and that these dealers are not better off improving upon the inside spread. This last condition requires that dealer m obtains *zero expected profit*, i.e., $S_b^* = \hat{S}(1,0)$. Third among the dealers who do not post the market spread, none should be better off pooling on the inside spread with dealer m. This imposes $\hat{S}(2,\Phi) > \hat{S}(1,0)$ or:

$$C(2,\Phi) > C(1,0).$$

We show that this is the case if $\Phi > \hat{\Phi}$. Since $C(M,.)$ increases with Φ, it is the case that $C(M,\Phi) > C(M,\hat{\Phi})$ when $\Phi > \hat{\Phi}(M,N)$. Furthermore $C(.,\Phi)$ decreases with M for $M \geq 2$. It follows that:

$$C(2,\Phi) \geq C(M,\Phi) > C(M,\hat{\Phi}) \quad for \quad \Phi > \hat{\Phi},$$

Now recall from the proof of Proposition 4 that $C(M,\hat{\Phi}) = C(1,0)$. Hence $C(2,\Phi) > C(1,0)$ for $\Phi > \hat{\Phi}$.

To sum up we have proved that the case in which only one dealer posts the inside spread $S_b^* = \hat{S}(1,0)$ and all the other dealers post a spread slightly greater than this spread is an equilibrium. Note that $S_b^* = \hat{S}(1,0)$ is the unique possible value for the inside spread in this case.□

Proof of Proposition 6.

Immediate using Equation (14).□

Proof of Proposition 7.

Consider equilibria in which all the dealers post the inside spread. Recall that the set of equilibrium spreads is non empty iff Inequality (22) holds true. Let:

$$\bar{x}^s(M) = \frac{C(1,0)x^l(M)}{C(M,0)},$$

be the value of $x^s(M)$ such that Inequality (22) is binding. Note that: $\bar{x}^s(M) > x^l(M)$ because $C(1,0) > C(M,0)$. Note also that $\bar{x}^s(M) < 1$ because $C(1,0) < MC(M,0), \forall M$. Furthermore note that the R.H.S of Inequality (22) increases with $x^s(M)$. Therefore Inequality (22) is satisfied **iff** $x^s(M) \leq \bar{x}^s(M)$.□

Proof of Proposition 8.

If $x^s(M) = 100\%, \forall M$ then $x^s(M) > \bar{x}^s(M)$ since $\bar{x}^s(M) < 1, \forall M > 1$ (see the proof of Proposition 6). This means that there is no equilibrium with more than one dealer posting the inside spread in this case. The case in which a single dealer posts a spread equal to $\hat{S}(1,0)$ is an equilibrium. The proof is similar to the proof of Proposition 5. In particular, the condition:

$$2C(2,0) < C(1,0),$$

is satisfied. It means that $\hat{S}(2,0) > \hat{S}(1,0)$ or no sidelined dealer can profitably match the quotes of the dealer posting the inside spread. Note that in this case, the dealer posting the spread earns zero profit.□

Proof of Corollary 1. Immediate using Lemma 2.□

Proof of Corollary 2.

From Proposition 2, we obtain:

$$\lambda^*(\Phi) = \sqrt{\frac{N\alpha Q(\sigma - S_b^*)}{cM^2(1+N-\Phi)^2}} \quad for \quad \Phi \leq \hat{\Phi}$$

Hence, in the zero expected profit equilibrium, the monitoring level of a dealer is:

$$\lambda^*(\Phi) = \sqrt{\frac{N(\alpha(1-\alpha)\sigma Q\beta\delta)}{cM^2[\alpha C(M,\Phi) + (1-\alpha)\beta\delta](1+N-\Phi)^2}} \quad for \quad \Phi \leq \hat{\Phi}.$$

Using the definition of $C(M,\Phi)$, this can be written as:

$$\lambda^*(\Phi) = \sqrt{\frac{N(\alpha(1-\alpha)\sigma Q\beta\delta)}{cM^2\left[\alpha\left(N(1+N-\Phi) + \frac{N}{2M_b}\right) + ((1-\alpha)\beta\delta)(1+N-\Phi)^2\right]}}.$$

It follows that $\frac{\partial\lambda^*}{\partial\Phi} > 0$ in this case. In the Pareto Dominant equilibrium, we obtain the same expression for λ^*, but $C(M,\Phi)$ is replaced by $\bar{C}(M,\Phi)$. As $\bar{C}(M,\Phi)$ decreases with Φ, it is direct that dealers' monitoring level increases with Φ. Thus, independently of the equilibrium we consider in the quoting stage, we obtain:

$$\lambda^*(0) < \lambda^*(\Phi) \quad \forall\Phi \leq \hat{\Phi}.$$

This proves the result.□

Proof of Corollary 3.

Recall that $M_b^* = M$ if $\Phi \leq \hat{\Phi}$. Using Proposition 2, we obtain that the aggregate monitoring level is:

$$\lambda_A^*(\Phi) + \gamma_A^*(\Phi) = M\lambda^* + N\gamma^* = (\Upsilon + N)\gamma^*.$$

which yields:

$$\lambda_A^*(\Phi) + \gamma_A^*(\Phi) = \sqrt{\frac{N\alpha Q(\sigma - S_b^*)}{c}}.$$

Consider the zero expected profit equilibrium. In this case, $S_b^* = \hat{S}(M,\Phi)$. As $\hat{S}(M,\Phi)$ increases with Φ, it follows from the previous equation that:

$$\lambda_A^*(\Phi) + \gamma_A^*(\Phi) < \lambda_A^*(0) + \gamma_A^*(0) \quad \forall\Phi \in (0,\hat{\Phi}].$$

Now consider the Pareto-Dominant equilibrium. In this case, $S_b^* = \bar{S}(\Phi)$. As $\bar{S}(M,\Phi)$ decreases with Φ, we now obtain that:

$$\lambda_A^*(\Phi) + \gamma_A^*(\Phi) > \lambda_A^*(0) + \gamma_A^*(0) \quad \forall\Phi \in (0,\hat{\Phi}].\square$$

Proof of Proposition 9.

Case 1: $\Phi \leq \hat{\Phi}$.

Computations yield

$$\frac{\partial C(M,\Phi)}{\partial N} = \frac{(N+1-\Phi)[(1-\Phi)2M+1]-2N}{2M(1+N-\Phi)^3}.$$

As $\Phi \leq \hat{\Phi} < \frac{1}{2}$, we obtain $\frac{\partial C}{\partial N} > 0$ which implies that $\frac{\partial \hat{S}}{\partial N} > 0$. Furthermore, we obtain:

$$\frac{\partial \bar{C}(M,\Phi)}{\partial N} = \frac{\frac{M(N+3)}{2(N+1)^3} - \frac{\partial C(M,\Phi)}{\partial N}}{M-1}.$$

Computations show that $\frac{\partial^2 C(M,\Phi)}{\partial \Phi \partial N} < 0$. It follows that $\frac{\partial^2 \bar{C}(M,\Phi)}{\partial \Phi \partial N} > 0$. Now for $\Phi = 0$, we get:

$$\bar{C}(M,0) = \frac{N}{N+1} + \frac{(M+1)N}{2M(N+1)^2},$$

which increases with N. This finally yields (since $\frac{\partial^2 \bar{C}(M,\Phi)}{\partial \Phi \partial N} > 0$):

$$\frac{\partial \bar{C}(M,\Phi)}{\partial N} > \frac{\partial \bar{C}(M,0)}{\partial N} > 0.$$

It follows that $\bar{S}(\Phi)$ increases with N.

Case 2: $\Phi > \hat{\Phi}$. In this case the equilibrium spread is:

$$\hat{S}(1,0) = \alpha\sigma\left(\frac{C(1,0)}{\alpha C(1,0) + (1-\alpha)\beta\delta)}\right) \tag{31}$$

$C(1,0)$ increases with N which implies that $\hat{S}(1,0)$ increases with N.□

Proof of Proposition 10.

Immediate using Equation (25).□

Proof of Corollary 4.

Consider an increase in Q. It shifts speculators' net expected profit upward for a given value of N (see Equation (25)). This induces entry of more speculators. The effect of σ is identical. The second part of the proposition follows directly from inspection of Equation (24).□

Proof of Proposition 11.

1st part. We sign the total impact of a change in Q on the equilibrium number of speculators. Let $\Pi_s^*(Q, N^*)$ be the net expected profit of speculators in equilibrium (the L.H.S of Equation (26)). We obtain:

$$\frac{dN^*}{dQ} = -\frac{\frac{\partial \Pi_s^*}{\partial Q}}{\frac{\partial \Pi_s^*}{\partial N^*}}.$$

It is straightforward that $\frac{\partial \Pi_s^*}{\partial Q} > 0$ and that $\frac{\partial \Pi_s^*}{\partial N^*} < 0$. It follows that $\frac{dN^*}{dQ} > 0$. Since $\hat{S}$ increases with the number of speculators, we conclude that $\hat{S}$ increases with Q.

2nd part. Using Proposition 2, we obtain that:

$$\lambda_A^* + \gamma_A^* = \Upsilon\gamma^* + N^*\gamma^* = \sqrt{\frac{\alpha N^* Q(\sigma - S_b^*)}{c}}$$

The number of speculators in equilibrium is such that each speculator's expected profit is zero in equilibrium. Hence, using Equation (25), we obtain:

$$\alpha Q(\sigma - S_b^*) = K\left[\frac{4N^*(N^* + 1 - \Phi)^2}{2N^*(N^* + 1 - \Phi) - (N^* - \Phi)^2}\right]$$

Thus we obtain:

$$\lambda_A^* + \gamma_A^* = \sqrt{\frac{K}{c}}\sqrt{\left[\frac{4(N^*)^2(N^* + 1 - \Phi)^2}{2N^*(N^* + 1 - \Phi) - (N^* - \Phi)^2}\right]}$$

It turns out that the term in bracket increases with N^*. Consequently $\lambda_A^* + \gamma_A^*$ increases with N^*. As N^* increases with Q, the proposition is proved.□

Proof of Proposition 12.

The inside spread in the Pareto-Dominant equilibrium increases with $\bar{C}(M, \Phi)$, which is given in Equation (20). Recall that we assume $x^s(M) = x^l(M) = 1/M$. Computations yield:

$$\frac{\partial \bar{C}(M, \Phi)}{\partial M} = \frac{1}{(M-1)^2}\left[C(M, \Phi) - C(1, 0) + \frac{N(M-1)}{2M^2(1 + N - \Phi)^2}\right]$$

The term in bracket increases with Φ. It is strictly negative for $\Phi = 0$ and strictly positive for $\Phi = \hat{\Phi}$ (because $C(1, 0) = C(M, \hat{\Phi})$). Thus there exists $\Phi^* \in (0, \hat{\Phi})$ such that $\frac{\partial \bar{C}(M, \Phi^*)}{\partial M} = 0$. For $\Phi < \Phi^*$, $\frac{\partial \bar{C}(M, \Phi)}{\partial M} < 0$ and for $\Phi > \Phi^*$, $\frac{\partial \bar{C}(M, \Phi)}{\partial M} > 0$.□

References

[1] Barclay, M., Christie, W., Harris, J., Kandel, E., and Schultz, P., 1999, The Effect of Market Reform on the Trading Costs and Depths of Nasdaq Stocks, *Journal of Finance.*, 54, 1-34.

[2] Battalio, R., Hatch, B., and Jennings, R., 1997, SOES Trading and Market Volatility, *Journal of Financial and Quantitative Analysis,* 32, 225-238.

[3] Benston, G., and Wood, R., 1998, Day Trading on Nasdaq's Automatic Small Order Execution System (SOES): Adverse Selection and Spreads, working paper, University of Memphis.

[4] Copeland, T., and Galai, D., 1983, Information Effects on the Bid-Ask Spread, *Journal of Finance,* 38, 1457-1469.

[5] Dennert, J., 1993, Price Competition between Market Makers, *Review of Economic Studies*, 60, 735-751.

[6] Fudenberg, D., and Tirole, J., 1991, Game Theory, MIT Press.

[7] General Accounting Office, 1998, The Effects of SOES on the Nasdaq Market, United States General Accounting Office Report 98-194.

[8] Grossman, S., Miller, M., Fishel, D., Cone,K. and Ross,D., 1995, Clustering and competition in asset markets. Working paper, Lexecon Inc., Chicago, IL.

[9] Harris, J., and Schultz, P., 1997, The Importance of Firm Quotes and Rapid Executions: Evidence from the January 1994 SOES rules change, *Journal of Financial Economics,* 45, 135-166.

[10] Harris, J., and Schultz, P., 1998, The Trading Profits of SOES Bandits, *Journal of Financial Economics,* 50, 39-62.

[11] Harris, J., 1994, Minimum Price Variations, Discrete Bid-Ask Spreads, and Quotations Sizes, *Review of Financial Studies,* 7, 149-178.

[12] Hinden, S., 1994, Nasdaq's Big Guns Send Trading Bandits Packing, Washington Post, February 7, 1994.

[13] Houtkin, H., 1998, Secrets of the SOES Bandit, McGraw-Hill.

[14] Kandel, E., and Marx, L., 1999, Odd-eight Avoidance as a Defense Against SOES Bandits, *Journal of Financial Economics.*, 51, 85-102.

[15] Kandel, E., and Marx, L., 1997, Nasdaq Market Structure and Spread Patterns, *Journal of Financial Economics*, 35, 61-90.

[16] Smith, J., Selway, J. and McCormick, T., 1998, The Nasdaq Stock Market: Historical Background and Current Operation, NASD Working Paper, 98-01.

[17] Stoll, H., 1992, Principles of trading market structure, *Journal of Financial Services Research*, 6, 75-107.

[18] Whitcomb, D., 1998, The NASDAQ Small Order Execution System: Myth and Reality, testimony before the House Committee on Commerce, Subcommittee on Finance, August 3, 1998.

Figure 1: Timing of the trading game

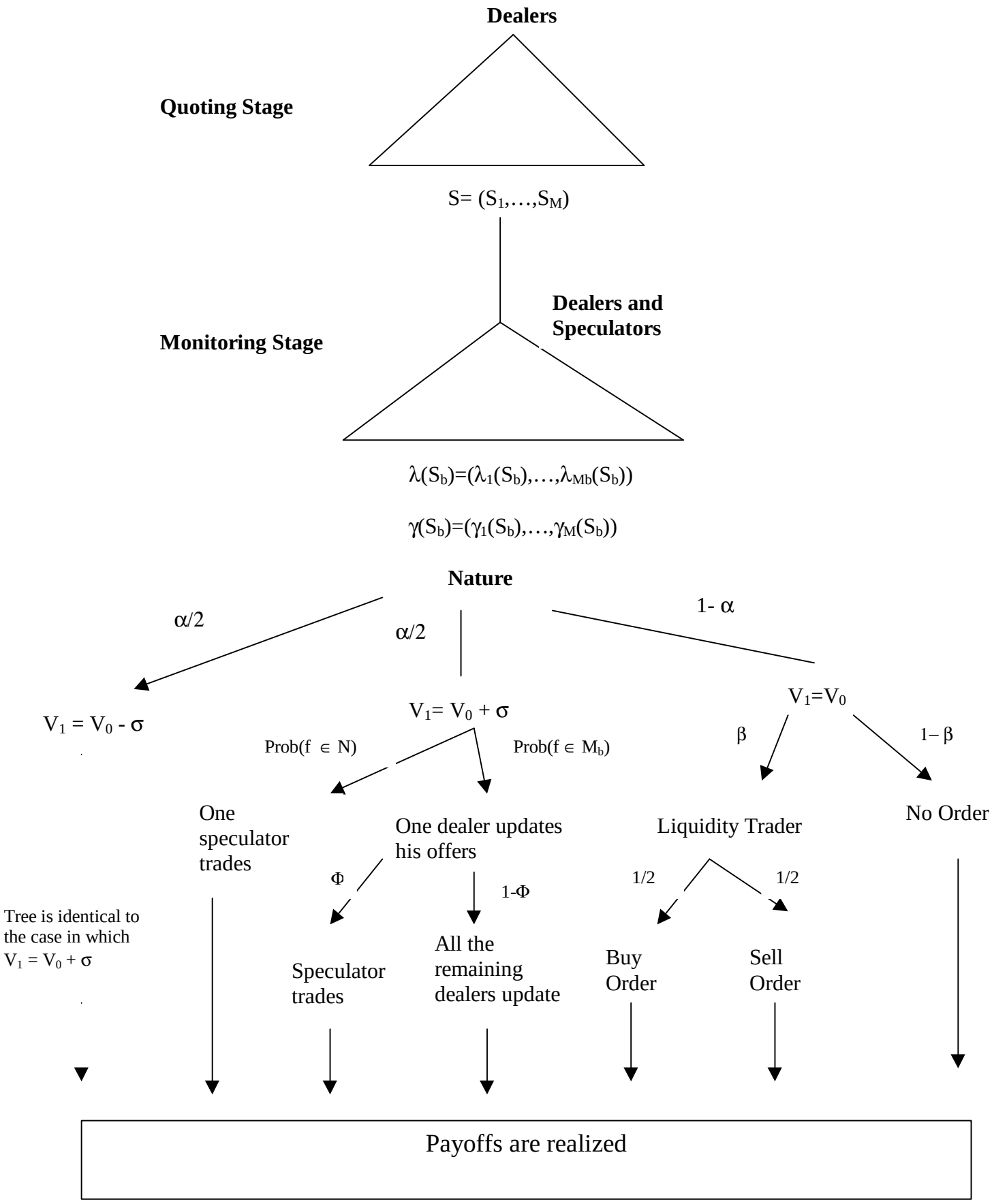
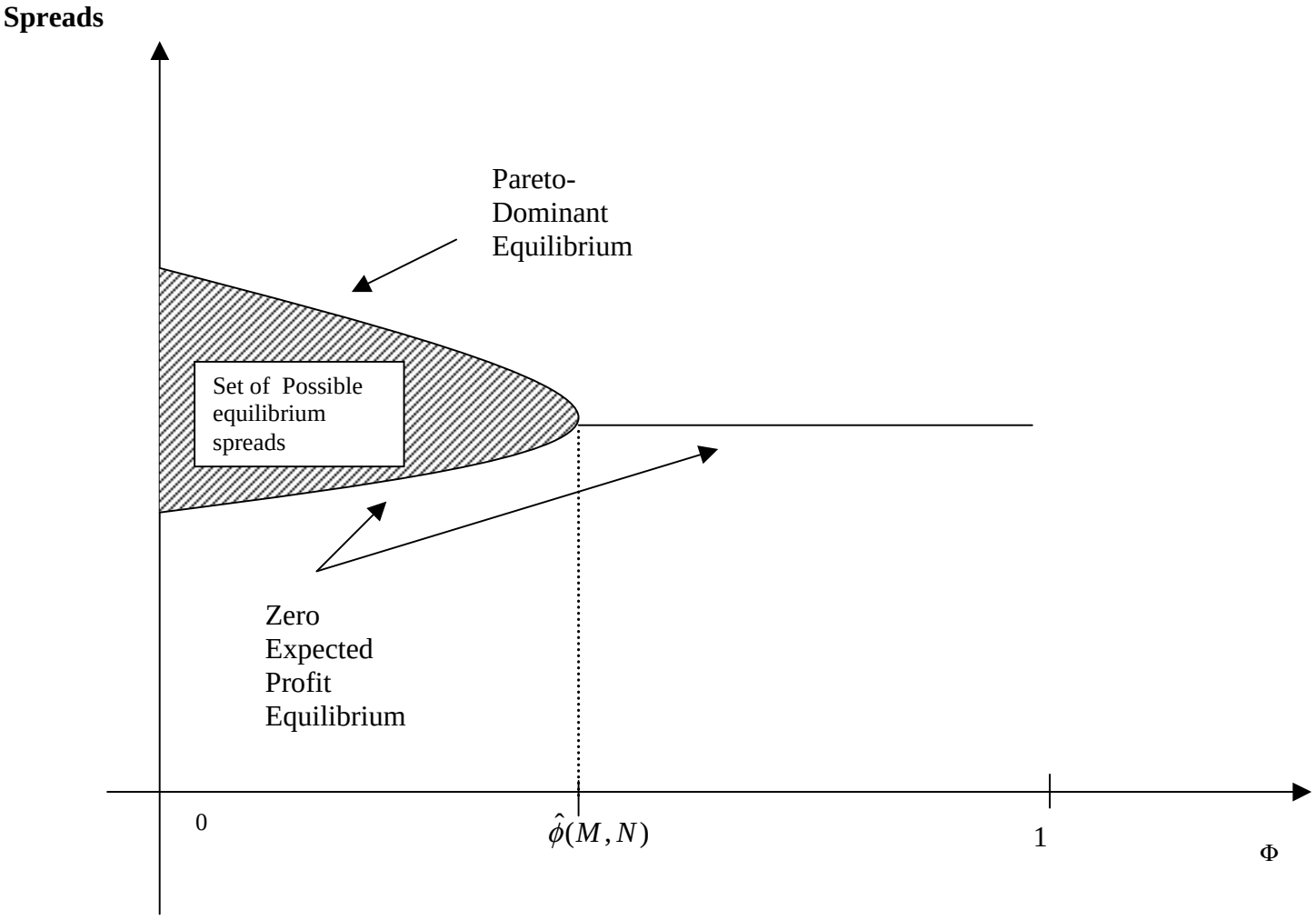


Figure 2: Equilibrium Spreads



Event	Action	Probability	Dealer i's Payoff
Liquidity Trader	Submits Buy Order	$\frac{(1-\alpha)\beta}{2}$	$x^l(M_b)\delta Q\frac{S_b}{2} - \Psi_d(\lambda_i)$
	Submits Sell Order	$\frac{(1-\alpha)\beta}{2}$	$x^l(M_b)\delta Q\frac{S_b}{2} - \Psi_d(\lambda_i)$
Good News	A Speculator Observes News First	$\frac{\alpha}{2}Prob(f \in \mathcal{N})$	$-x^s(M_b)Q\frac{(\sigma-S_b)}{2} - \Psi_d(\lambda_i)$
$(v_1 = v_0 + \frac{\sigma}{2})$	Dealer i Observes News First	$\frac{\alpha}{2}Prob(f = i)$	$-\Psi_d(\lambda_i)$
	Dealer k, $k \neq i$ Observes News First and Speculator is Second	$\frac{\alpha}{2}\Phi Prob(f \in \mathcal{M}_b \backslash i)$	$-x^s(M_b - 1)Q\frac{(\sigma-S_b)}{2} - \Psi_d(\lambda_i)$
	Dealer k, $k \neq i$ Observes News First and Dealers Follow	$(1-\Phi)\frac{\alpha}{2}Prob(f \in \mathcal{M}_b \backslash i)$	$-\Psi_d(\lambda_i)$
Bad News	A Speculator Observes News First	$\frac{\alpha}{2}Prob(f \in \mathcal{N})$	$-x^s(M_b)Q\frac{(\sigma-S_b)}{2} - \Psi_d(\lambda_i)$
$(v_1 = v_0 - \frac{\sigma}{2})$	Dealer i Observes News First	$\frac{\alpha}{2}Prob(f = i)$	$-\Psi_d(\lambda_i)$
	Dealer k, $k \neq i$ Observes News First and Speculator is Second	$\frac{\alpha}{2}\Phi Prob(f \in \mathcal{M}_b \backslash i)$	$-x^s(M_b - 1)Q\frac{(\sigma-S_b)}{2} - \Psi_d(\lambda_i)$
	Dealer k, $k \neq i$ Observes News First and Dealers Follow	$(1-\Phi)\frac{\alpha}{2}Prob(f \in \mathcal{M}_\lfloor \backslash i)$	$-\Psi_d(\lambda_i)$
No Order & No Information		$(1-\beta)(1-\alpha)$	$-\Psi_d(\lambda_i)$

Table 1: Dealer i's Payoffs

Event	Action	Probability	Speculator j's Payoff
Liquidity Trader	Submits Buy Order	$\frac{(1-\alpha)\beta}{2}$	$-\Psi_s(\gamma_j)$
	Submits Sell Order	$\frac{(1-\alpha)\beta}{2}$	$-\Psi_s(\gamma_j)$
Good News	Speculator j observes the news first	$\frac{\alpha}{2}Prob(f=j)$	$x^s(M_b)M_bQ\frac{(\sigma-S_b)}{2}-\Psi_s(\gamma_j)$
$(v_1=v_0+\frac{\sigma}{2})$	Speculator k, $k\neq j$, is first to observe news	$\frac{\alpha}{2}Prob(f\in\mathcal{N}\backslash j)$	$-\Psi_s(\gamma_j)$
	Dealer Updates Quote and Speculator j Reacts First	$\frac{\alpha}{2N}\Phi Prob(f\in\mathcal{M}_b)$	$x^s(M_b-1)M_bQ\frac{(\sigma-S_b)}{2}-\Psi_s(\gamma_j)$
	Dealer Updates Quotes and Speculator j is Not First to React	$(1-\Phi)\frac{\alpha}{2}Prob(f\in\mathcal{M}_b)$	$-\Psi_s(\gamma_j)$
Bad News	Speculator j observes the news first	$\frac{\alpha}{2}Prob(f=j)$	$x^s(M_b)M_bQ\frac{(\sigma-S_b)}{2}-\Psi_s(\gamma_j)$
$(v_1=v_0-\frac{\sigma}{2})$	Speculator k, $k\neq j$, is first to observe news	$\frac{\alpha}{2}Prob(f\in\mathcal{N}\backslash j)$	$-\Psi_s(\gamma_j)$
	Dealer Updates Quote and Speculator j Reacts First	$\frac{\alpha}{2N}\Phi Prob(f\in\mathcal{M}_b)$	$x^s(M_b-1)M_bQ\frac{(\sigma-S_b)}{2}-\Psi_s(\gamma_j)$
	Dealer Updates Quotes and Speculator j is Not First to React	$(1-\Phi)\frac{\alpha}{2}Prob(f\in\mathcal{M}_b)$	$-\Psi_s(\gamma_j)$
No Order & No Information		$(1-\beta)(1-\alpha)$	$-\Psi_s(\gamma_j)$

Table 2: Speculator j's Payoffs

Table 3: Summary Statistics

Variable	Mean	Median	Std. Dev.	Minimum	Maximum
Number of SOES *clusters*	204	61	540	0	6134
Total Number of SOES trades	2466	1047	5651	50	62178
Total Number of non-SOES trades	9998	5441	16470	924	151236
Bid-ask spread	1.30%	1.14%	0.68%	0.11%	3.97%
Volatility	0.87%	0.86%	0.34%	0.16%	2.68%
Maximum SOES quantity	968.39	1000	139.70	200	1000
Number of Dealers	22.28	20.52	10.25	5.33	63.52
Average trade size	1666	1483	700	595	5381
Market capitalization	2457	791	8885	71	107500
Average Price	26.71	23.53	17.27	5.02	130.94

Table 3: The mean, median, standard deviation, minimum, and maximum for variables in our sample of 310 stocks are reported. A SOES *cluster* is defined as an uninterrupted sequence of three SOES trades of maximum size, at the same price, within 30 seconds. The total number of SOES trades includes all trades through the SOES system. The total number of non-SOES trades include all trades during regular trading hours that were not submitted through the SOES system. The bid-ask spread is measured as the time-weighted average of the relative inside spread. Volatility is measured by the standard deviation of the half-hour returns computed based on the mid-quotes. The maximum SOES size is a discrete variable with values equal to 1000, 500, or 200 shares. The number of dealers is computed as an time-series average of the number of active dealer in each stock. The average trade size is measured as the average number of shares traded in trades that are not part of a SOES *cluster*. The market capitalization and the average price are based on monthly CRSP data.

Table 4: Correlation Matrix

	soes	sprd	vlty	maxQ	Ndlr	liqD	mkcp
sprd	-0.6835						
vlty	-0.2344	0.4871					
maxQ	0.2813	-0.0747	-0.1184				
Ndlr	0.1081	-0.2568	-0.3185	0.1882			
liqD	-0.1850	0.0438	-0.1262	-0.7325	-0.1707		
mkcp	0.4636	-0.7509	-0.5240	0.0792	0.4249	-0.0074	
avgP	0.5923	-0.7572	-0.3415	0.0569	-0.1691	-0.0225	0.6966

Table 4: The variables in the correlation matrix are defined as: the log odds ratio of the probability of a SOES *cluster* (soes), the average time-weighted bid-ask spread (sprd), the maximum SOES quantity (maxQ), the number of dealers (Ndlr), the average trade size relative to the maximum SOES quantity (liqD), the logarithm of the market capitalization (mkcp), the logarithm of the average price (avgP).

Table 5: Estimation Results

	SOES Equation		Spread Equation	
	Coefficient	P-value	Coefficient	P-value
Constant	-2.528431	<0.001	0.0592633	<0.001
SOES proxy			0.0043574	0.079
Bid-ask spread	-70.85455	<0.001		
Volatility	31.93144	<0.001	0.1784688	0.058
Maximum SOES quantity	0.0010533	<0.001		
Number of dealers			-0.0003185	<0.001
Liquidity Demand			0.0001834	0.453
Market capitalization			0.0002335	0.686
Average price			-0.0109728	<0.001

Table 5: The parameter estimates with p-values for the system given in Equation (**??**), adding market capitalization and the average price are reported. The system was estimated using three-stage least squares. The SOES proxy is the log odds ratio for the probability of a SOES *cluster*, the bid-ask spread is the time-weighted inside spread, the volatility is defined as the standard deviation of half-hour mid-quote returns, the maximum SOES quantity is either 1000, 500, or 200, the number of dealers is a time-series average of the number of active dealers, the liquidity demand is the average size of trades that are not part of a *cluster* divided by the maximum SOES quantity. The logarithm of the market capitalization and the average price are used in the Spread Equation.

Table 6: Reduced Form Equations

	SOES Equation		Spread Equation	
	Coefficient	P-value	Coefficient	P-value
Constant	-5.196232	<0.001	0.0368048	<0.001
Volatility	13.36749	0.259	0.2412264	0.012
Maximum SOES quantity	0.0011604	<0.001	0.00000389	0.058
Number of dealers	0.0200559	<0.001	-0.0002404	<0.001
Liquidity Demand	0.0444787	0.138	0.0001979	0.333
Market capitalization	-.1212504	0.006	0.000063	0.857
Average price	0.7711062	<0.001	-0.008196	<0.001
R-squared	0.4570		0.7375	
F(6, 303)	54.36		99.99	

Table 6: The parameter estimates, with p-values in parenthesis, are reported for the reduced form equations that correspond to the system of simultaneous equations of Table 5. The SOES proxy is the log odds ratio for the probability of a SOES *cluster*, the bid-ask spread is the time-weighted inside spread, the volatility is defined as the standard deviation of half-hour mid-quote returns, the maximum SOES quantity is either 1000, 500, or 200, the number of dealers is a time-series average of the number of active dealers, the liquidity demand is the average size of trades that are not part of a *cluster* divided by the maximum SOES quantity. The logarithm of the market capitalization and the average price are used in the Spread Equation.

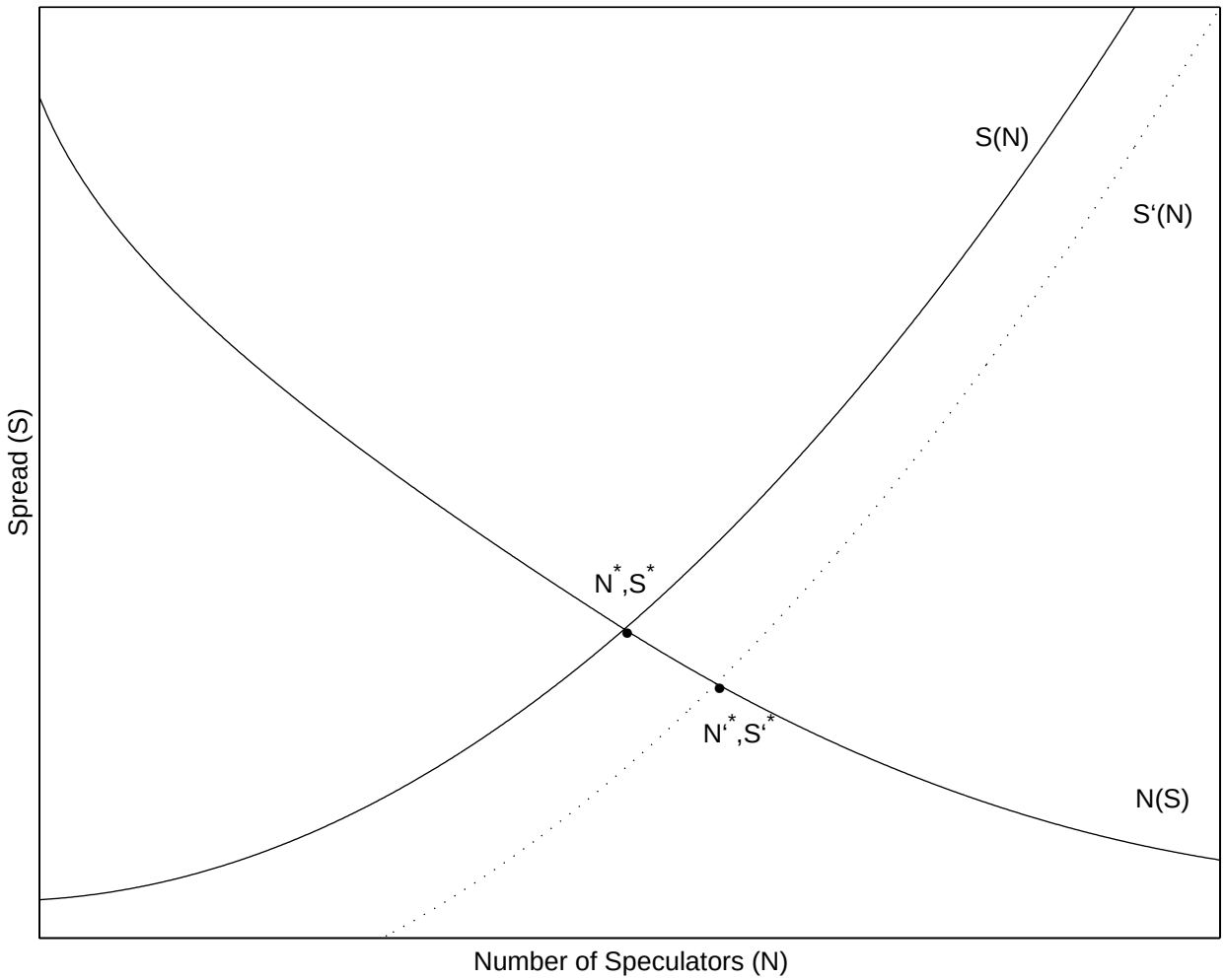


Figure 3: This graph qualitatively illustrates the simultaneous determination of the bid-ask spread and the number of speculators. The $S(N)$ curve represents the spread as a function of the number of dealers. The $N(S)$ represent the number of dealers as a function of the spread. The initial equilibrium point is represented by N^* and S^*. The dotted line illustrates the qualitative effect of increasing the liquidity demand, i.e., an increase in δ. This change shifts the $S(N)$ curve out to $S'(N)$ and the direct effect is to lower the spread without changing the number of dealers. The indirect effect is that the number of speculators increase as a result of the lower spread and this results in a new equilibrium point N'^* and S'^*.

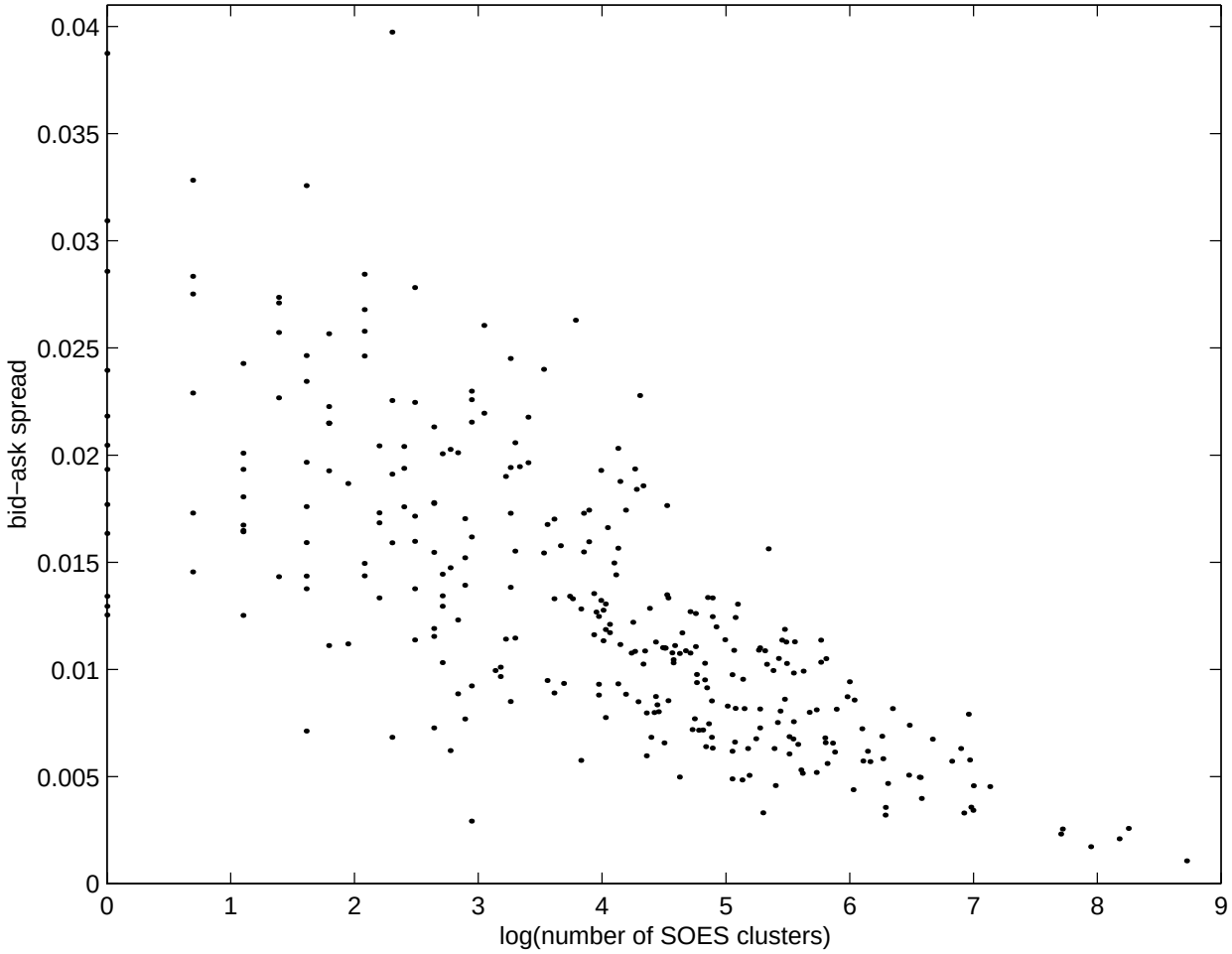


Figure 4: The log(number of SOES clusters) is plotted against the bid-ask spread. A SOES cluster is defined as 3 maximum size SOES trade submitted within 30 seconds at the same price.